Mechanisms of translational regulation in bacteria: Impact on codon usage and operon organization

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)
im Fach Biologie

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät I
der Humboldt-Universität zu Berlin

von
Herrn Diplom-Physiker Kajetan Bentele

Präsident der der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I:
Prof. Stefan Hecht PhD

Gutachter:
1. Prof. Dr. Markus Kollmann
2. Prof. Dr. Nils Blüthgen
3. Prof. Dr. Zoya Ignatova

Tag der mündlichen Prüfung: 16.05.2013

Ich widme diese Arbeit
meiner Familie und meinen Freunden

Abstract

Translation is the final step in the fundamental process of protein biosynthesis, the proper course of which is of utmost importance to the living cell. Here we investigate the relationship between translational efficiency and codon usage at the gene start. It is known for some organisms that usage of synonymous codons at the beginning of genes deviates from the codon usage elsewhere in the genome. By systematically analyzing about 400 bacterial genomes we find that this phenomenon is widespread but differs markedly in strength. We show that this deviation in codon usage is caused by the need to suppress RNA secondary structure around the translation start site, thereby allowing efficient initiation of translation. This pressure to reduce folding increases with the GC-content of the respective genome. In contrast to the current hypothesis that codon usage is adapted in order to slow down early elongation, we conclude that the observed enrichment of rare codons is a consequence of suppressing mRNA structure around the ribosome binding site (RBS). We validate this hypothesis experimentally by varying independently codon usage and folding of mRNA and measuring protein- and mRNA-levels.

We investigate further driving forces for genome organization by studying the impact of gene order within an operon on the fitness of bacterial cells. Operons group functionally related genes which are together transcribed as single mRNAs in *E. coli* and other bacteria. Correlation of protein levels is thus to a large extent attributed to this coupling on the transcriptional level. In addition, the initiation of ribosomes at the RBS of adjacent genes within an operon may be dependent on each other. Such translational coupling can further stabilize a desired stoichiometry between proteins. Here, we study the role of translational coupling in robustness of *E. coli* chemotaxis. We demonstrate experimentally translational coupling of chemotaxis genes and the beneficial effects of pairwise expression of genes from bicistronic constructs on chemotactic performance. By employing a model of translational coupling and simulating the underlying signal transduction network we show that certain permutations of genes contribute more to robustness of chemotaxis than others. We conclude that translational coupling is an important determinant of the gene order within the chemotaxis operon.

Both these findings show that requirements for efficient gene expression and robustness of cellular function have a pronounced impact on the genomic organization, influencing the local codon usage at the beginning of genes and the order of genes within operons.

Zusammenfassung

Die Translation ist der letzte Schritt der Proteinbiosynthese, ein Prozess, der von außerordentlicher Bedeutung für die Zelle ist. Hier untersuchen wir den Zusammenhang zwischen der Translationseffizienz von Genen und der Häufigkeit bestimmter Codons am Genanfang in bakteriellen Genomen. Für einige Organismen wurde gezeigt, dass die Häufigkeitsverteilung der Codons am Anfang der Gene eine andere ist als sonst im Genom. Durch die systematische Untersuchung von ungefähr 400 bakteriellen Genomen haben wir festgestellt, dass dieses Phänomen sehr weit verbreitet ist, sich jedoch in der Ausprägung zum Teil erheblich unterscheidet. Unsere Analyse zeigt, dass der Grund dieser Abweichung in der Notwendigkeit liegt, RNA Sekundärstruktur in der Nähe des Translationsstarts zu vermeiden. Der evolutionäre Druck die Faltung der RNA zu unterdrücken ist dabei umso stärker, je größer der GC-Gehalt des jeweiligen Genoms ausfällt. Unsere Ergebnisse stehen im Gegensatz zur gegenwärtigen Hypothese, wonach am Anfang von Genen solche Codons präferentiell benutzt werden, die eine Verlangsamung der Ribosomen in der frühen Elongationsphase zur Folge haben sollten. Dieser These zufolge führt das zu einer Anreicherung von seltenen Codons, wohingegen wir zu dem Schluss gekommen sind, dass dies nur eine Folge der Notwendigkeit ist, die Ribosombindestelle (RBS) einer RNA möglichst unstrukturiert zu belassen. Wir haben diese Hypothese experimentell validiert, indem wir den Gebrauch synonymer Codons unabhängig von der mRNA Faltung variiert und die Protein und mRNA Häufigkeit dieser Konstrukte bestimmt haben.

Im zweiten Teil der Arbeit untersuchen wir die Genomorganisation auf einer anderen Ebene: Den Einfluss der Genreihenfolge innerhalb eines Operons auf die Fitness von *E. coli*. In den Genomen von *E. coli* oder anderen Bakterien fasst ein Operon Gene zusammen, die in einer funktionellen Beziehung zueinander stehen und zusammen transkribiert werden. Die Korrelation zwischen den Häufigkeiten solcherart kodierter Proteine ist daher zu einem Teil auf die Kopplung der Transkription zurückzuführen. Hinzu kommt, dass die Initiation der Ribosomen an benachbarten Gene voneinander abhängen kann. Diese zusätzliche translationale Kopplung kann eine gewünschte Stöchiometrie zwischen Proteinen weiter stabilisieren. Hier haben wir die Rolle der translationalen Kopplung für die Robustheit des Chemotaxis Signalweges in *E. coli* untersucht. Wir haben experimentell gezeigt, dass es eine Kopplung auf der Ebene der Translation zwischen den Chemotaxis-Genen gibt und dass die paarweise Überexpression dieser Gene weitaus besser toleriert wird als die einzelner Gene. Mit Hilfe eines Modells für die translationale Kopplung sowie für den Chemotaxis Signalweg konnten wir zeigen, dass bestimmte Permutationen der Gene mehr zur Robustheit beitragen als andere. Die translationale Kopplung ist daher ein wichtiger Faktor, der die Anordnung der Gene innerhalb des Chemotaxis Operons bestimmt.

Diese Arbeit zeigt, dass die Anforderungen einer effizienten Genexpression sowie die Robustheit essentieller zellulärer Funktionen einen wichtigen Einfluss auf die Organisation eines Genoms haben können: Einerseits bei der Wahl der Codons am Anfang der Gene, andererseits auf die Ordnung der Gene innerhalb eines Operons.

Contents

1. Introduction

The genomes we nowadays observe are the result of a billion year long evolution (36). Each genome consists of one ore more large non-branched polymers (DNA) which encode the genetic information in the sequence of four different "letters", the so called nucleotides. Mutations constantly change genomes in a random fashion often having a deleterious effect on the organism (113). However, a mutation may also lead to a better adaption of an individual to its environment and thus increase its reproductive success. As the new trait has a genetic basis it is passed to the next generation and if it continues to be beneficial, the frequency of the genotype will increase and eventually take over the whole population. Such differential reproduction of genotypes is called selection (113). Additionally, stochastic fluctuations in the frequency of neutral genetic traits become important in small populations. Since such changes are random they are referred to as genetic drift (44, 113).

The interplay between mutations, causing genetic variations, selection and genetic drift, either leading to the increase or decrease in the frequency of genetic variants, shapes the genetic composition of a population. For a complete description, we would need the specification of the genome and spatial location of every individual at one instant in time (44). However, for traits which have been fixed in the population, a single sequenced genome may be considered as a representative of the evolutionary process which has shaped its overall structure and composition. Moreover, selective forces prevalent across the border of species should become apparent when we compare different genomes.

In this thesis we investigate the relation between the structure of bacterial genomes and the process of translation at two different organizational levels (fig. 1.1). We analyze the causes for differential codon usage at the beginning of genes as well as the question whether there are selective forces influencing the order of genes in an operon. We show first that the need to suppress compact RNA structure around the translation start site is an important determinant of the codon usage at the beginning of genes. To this end we made use of the comparative analysis of different genomes. Second, we demonstrate theoretically and experimentally that translational coupling between adjacent genes in a polycistronic mRNA is a crucial factor influencing the order of genes in the chemotaxis operon. Here we studied a single genome by employing mathematical models to simulate alternative scenarios of genomic organization.

Both aspects of genomic organization we investigate in this thesis are related to the translational process, the final step in one of the most important tasks cells have to accomplish:

The conversion of genetic information stored in the DNA as a sequence of nucleotides into functional proteins. The synthesis of proteins requires the production of messenger RNAs (mRNAs), short-lived copies of the genes. The nucleotide sequences of these mRNAs are then converted by huge protein-RNA complexes, the ribosomes, into the sequence of amino acids defining the primary structure of the proteins. Which amino acids are incorporated into the growing peptide chain is determined by triplets of nucleotides, so called codons. However, there are only 20 amino acids but 64 different triplets of nucleotides encoding them. Consequently, the genetic code is degenerate: Except for tryptophan and methionine, the amino acids are encoded by two, four or six different so called synonymous codons. Hence an organism can tune the codon sequence according to its needs without changing the amino acid sequence of the expressed protein. On a genome-wide scale not all synonymous codons are used with the same frequency: Some are preferred to others. This is termed codon usage bias, each organism having its specific one (49, 131, 54). For *E. coli* it is known that the abundance of individual transfer RNAs and the frequency of usage for the corresponding cognate codons are correlated (60, 34). Cellular levels of transfer RNAs are believed to be important in modulating the elongation rate of ribosomes along the transcript (169, 170). Interestingly there is an enrichment of rare codons at the beginning of genes suggesting different driving forces shaping codon usage than elsewhere in the gene (42, 154). It was hypothesized that rare codons may be preferentially used to reduce elongation speed at the beginning of a gene in order to reduce the likelihood of ribosomal "traffic jams" along the mRNA (154). Before ribosomes can start to elongate, however, they have to bind to mRNA thereby initiating translation. We expected that this necessity is also an important factor shaping the sequence near the translation start site of a gene. There is plenty of evidence that mRNA secondary structure around the translation start site is an important determinant of translation efficiency (30, 31, 76). Moreover, it was found that suppression of mRNA structure around the translation start site is prevalent in *E. coli* and a widespread phenomenon found across many different genomes (76, 50). This gives rise to the hypothesis that enrichment of rare codons is rather a byproduct of the selection for efficient translation initiation and in turn for suppression of mRNA structure. To investigate the relationship between structure formation and codon usage we analyzed around 400 bacterial genomes and found that deviation of codon usage in the first few codons is widespread but differs markedly in strength. We demonstrate that this deviation is more pronounced if the genome is GC-rich and thus folding energy of mRNA is large. To address the question whether there is a selective enrichment of rare codons, we looked more closely at the usage of the most abundant and the most rare codons at the beginning of genes. Consistently with the hypothesis that suppression of RNA structure around the translation start site drives codon usage, we found that codons which reduce GC-content were preferentially selected at gene start. Such local depletion of GC-content can destabilize mRNA secondary structures and in turn may allow for efficient ribosome binding to the mRNA, which is necessary for translation initiation. In

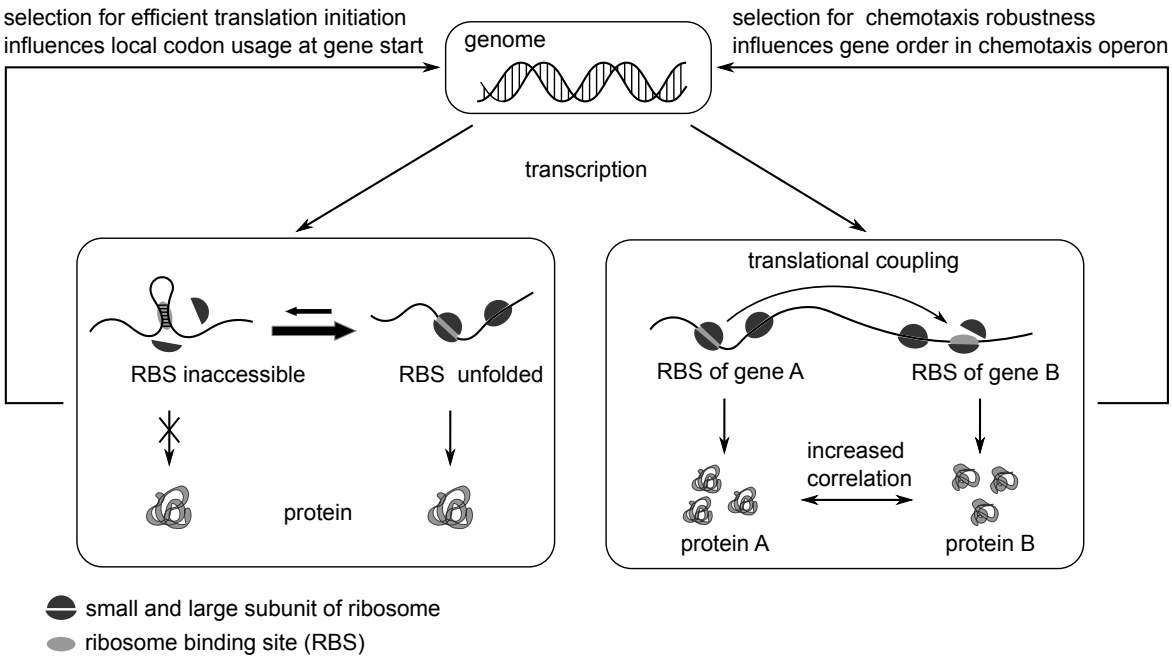


Figure 1.1.: Graphical abstract of the thesis. Genes are transcribed into mRNAs which are in turn translated into proteins. This thesis studies two aspects of the translational process in bacteria: Initiation of translation and coupling of translation between adjacent genes in a polycistronic mRNA.
Single stranded RNAs are capable of forming secondary structures. Such structures may render the ribosome binding site of a gene inaccessible and hence an evolutionary pressure to suppress structure formation around the translation start site is expected. Here, we investigated the impact of this selective force on the codon usage at the beginning of genes in *E. coli* and around 400 other bacterial genomes.
Many functional related genes in bacteria are combined in operons and transcribed as a single polycistronic mRNA, thus being strongly co-regulated on the transcriptional level. However, even translation may be coupled in these mRNAs leading to an even stronger correlation of expression levels. Here, we show experimentally the presence of translational coupling between chemotaxis gene pairs and investigate the impact coupling of selective genes may have on chemotactic performance and thus on gene order in the chemotaxis operons.

addition, we show experimentally that changing the folding energy while keeping the same codon usage at the beginning of native *E. coli* genes markedly affects translation efficiency. In contrast, alterations of the codon usage while maintaining the same folding energy led to less conclusive results. We therefore conclude that the enrichment of rare codons in bacterial genomes is most likely a consequence of the need to suppress mRNA structure around the ribosome binding site and not due to the selection of rare codons *per se* (fig. 1.1).

Genomes are organized on many levels with the sequence of nucleotides being the most fundamental one. At the next level, sequences of codons constitute protein-coding genes. In bacteria, many functional-related genes are in turn organized within larger groups, called operons (120). Such clustered genes underlie a common transcriptional control and as a

consequence are transcribed together in polycistronic mRNAs. Due to this transcriptional coupling protein levels are correlated. The proteins of the chemotaxis pathway in *E. coli* are expressed as two polycistronic units. In the second part of the thesis we address the question whether in addition to the clustering also the order of genes within these operons is under selective pressure or just the outcome of chance.

The chemotaxis system as a whole is under strong selection as it enables bacteria to search for optimal growth conditions thereby conferring a competitive advantage. Cells are directed towards favorable environments by a biased random walk (12, 1). This mechanism is implemented as a network of interacting proteins, translating extracellular signals on a fast timescale into phosphorylation of response regulators, while assuring adaption to external stimuli on a slower timescale. Precise adaptation to a wide range of stimulus strengths emerges from the topological properties of the network and does not rely on a fine tuning of parameters (7, 3). The topology does not only assure precise adaptation but also robustness of the pathway output against correlated fluctuations of the pathway components (73). In light of these findings, it becomes apparent that clustering of chemotaxis genes in two operons is the strategy selected by evolution to assure the correlation of protein levels. However, the reason for the order of genes within the operons remained unresolved. We investigated whether translational coupling, defined as the interdependence of translational efficiency between neighboring genes within a single polycistronic mRNA, can be responsible for the observed gene order. Translational coupling was previously described in *E. coli* and can stabilize a desired stoichiometry between proteins expressed from the same operon (109, 8, 85). We show experimentally translational coupling for most pairs of chemotaxis genes and confirm that coexpression of these pairs improves chemotactic performance compared to overexpression of single genes. To demonstrate the benefit of translational coupling on chemotactic performance we modeled its impact on the robustness of signal processing. This required to simulate the underlying signal transduction network using a model with ordinary differential equations based on the law of mass action. Thereby we can show that robustness of the pathway against the uncorrelated variations in protein levels can be enhanced by a selective pairwise coupling of chemotaxis genes. Furthermore, we demonstrate that the observed order of genes in *E. coli* ranks among the best in terms of noise compensation. In addition, we develop arguments independent of model details corroborating the importance of pairwise coupling. The order of genes in the chemotaxis operon may therefore be influenced by the need to pair specific genes which are then translationally coupled in turn reducing the negative effects of uncorrelated noise on the pathway.

This thesis is structured as follows. The next chapter provides background information about gene expression in bacteria, covering transcription and translation as well as gene expression noise. In addition we develop a coarse grained mathematical model to describe the translation process. In chapter 3 we show how suppression of mRNA structure shapes codon usage at the gene start in bacteria. We present the results of our bioinformatics

analysis, evolutionary simulations and experimental findings corroborating our hypothesis that suppression of mRNA around the translation start is the main driving force for the observed codon usage at beginning of genes. The mathematical model developed in chapter 2 is used to understand the possibly harmful effects of slowly translated codons at the gene start on the translation efficiency. Chapter 4 deals with the impact of translational coupling on the gene order within the chemotaxis operon. The presented experimental findings are accompanied by a theoretical analysis of the relation between translational coupling and selection for chemotaxis robustness. Our mathematical model of the translation process serves as a starting point to develop a framework for modeling translational coupling between adjacent genes in an operon. Finally, chapter 5 concludes the thesis. There we discuss the significance of our results and suggest possible future experiments building on our findings.

2. Gene expression in bacteria

2.1. The central dogma of molecular biology

Cells need numerous different proteins to guarantee their survival and proliferation. These proteins are necessary for maintaining or changing the structure, catalyzing metabolic reactions, driving transport, processing signals, regulating cellular processes, organizing cell replication, and finally for building up all the proteins themselves. The information required to construct all proteins is stored in the deoxyribonucleic acid (DNA) of a cell. Stretches of DNA, called genes, function as templates for the synthesis of functional gene products. The set of reactions controlling the abundance of these gene products is called gene expression (117). Often the final product of gene expression is a protein and the corresponding genes are therefore referred to as protein-coding genes. In addition there are also non-protein coding genes, whose products are functional ribonucleic acid (RNA) molecules, including ribosomal RNA and transfer RNA. In protein synthesis, RNA functions as an intermediate product and is therefore called messenger RNA (mRNA) (87, 113).

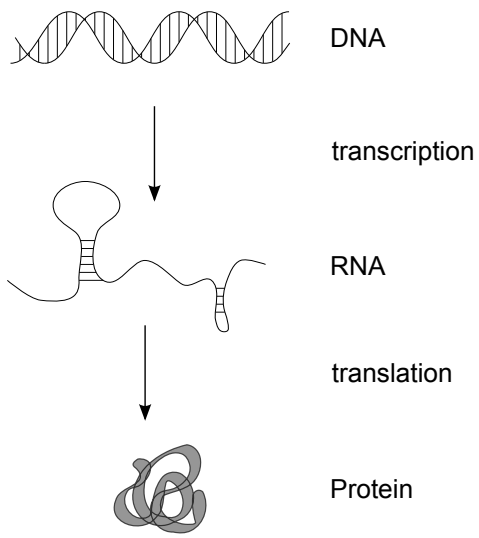


Figure 2.1.: Protein biosynthesis

Gene expression is a multi-step process (fig. 2.1). The point of origin is the cellular DNA, a double-stranded polymer built up from four different nucleotides. The sequence of these four "letters" encodes the information necessary to synthesize functional RNA and proteins. An enzyme, RNA polymerase, stepwise links together single nucleotides complementary to the template DNA strand, resulting in either precursor of functional RNA or mRNA. The produced RNA is thus a copy of the nucleotide sequence of the other DNA strand and the whole process of copying is called transcription. Proteins are also polymers build up from 20 different amino acids. Thus the ribonucleotide sequence has to be converted into a string of amino acids, forming the primary structure of the protein. This process, referred to as translation, is catalyzed by ribosomes, macromolecular complexes consisting of ribosomal RNA (rRNA) and proteins. Triplets of nucleotides in the mRNA, so called codons, determine which of the 20 amino acids

is chosen next. Proteins fold during and after translation into a definite three dimensional structure in order to function properly. Together with the replication of DNA, the flow of sequential information from DNA to proteins in these processes is called the central dogma of molecular biology (25).

Several goals have to be met by the expression machinery to maintain proper cell function: First, RNAs and proteins have to be assembled without errors. Second, cells face a trade-off between energy efficiency and flexibility (77). RNA and protein synthesis consumes a significant amount of cellular energy resources (163). Thus proteins and mRNAs should be stable in order to achieve best energy efficiency. However this might lead to a loss of flexibility, since levels of stable mRNAs and proteins cannot be adapted rapidly to a changing environment. Hence a regulation of gene expression is necessary to avoid wasting of cellular resources, while being responsive to a fluctuating environment. As an example, consider the change of the carbon source from glucose to lactose in a medium containing *Escherichia coli* (*E. coli*) cells. *E. coli* preferentially digests glucose, however, if no glucose is present but only lactose, the bacterium has to adapt to the new environment. This is accomplished by expressing the *lac* genes coding for enzymes which are necessary to metabolize the lactose. Thus *E. coli* can adapt to these new conditions (115, 47).

We can describe the process of gene expression and its regulation by a simple mathematical model comprising the two steps of transcription and translation. RNA polymerases synthesize mRNA from DNA with a rate k_m. The transcription rate $k_m = k_m(s, r)$ is a function of the integrated signals s and the concentration r of available RNA polymerases. The degradation of mRNA is modeled as a unimolecular reaction with the rate constant γ_m

$$\emptyset \xrightarrow{k_m} \mathrm{m} \xrightarrow{\gamma_m} \emptyset.$$

Ribosomes process mRNA and translate it into proteins with a rate constant $k_p = k_p(R)$, which is a function of free ribosome concentration R. As in the case of mRNA, proteins are assumed to degrade in a first order reaction with rate constant γ_p,

$$\mathrm{m} \xrightarrow{k_p} \mathrm{p} \xrightarrow{\gamma_p} \emptyset.$$

The rate equations for these processes therefore read

$$\frac{d}{dt}m = k_m - \gamma_m m \tag{2.1}$$

$$\frac{d}{dt}p = k_p m - \gamma_p p, \tag{2.2}$$

where we denoted the concentration of the different species in italic letters. For the sake of

parameter	value	gene with 500 codons
k_m	$\lesssim 80$ bp s^{-1}	$\lesssim 0.05$ mRNA s^{-1}
γ_m	$\sim (3-8) \times 10^{-3}$ s^{-1}	
k_p	$\lesssim 40$ aa s^{-1}	$\lesssim 0.08$ protein mRNA^{-1} s^{-1}
γ_p	$\sim 6 \times 10^{-4}$ s^{-1}	

Table 2.1.: Typical parameter values for gene expression in bacteria (1). The degradation rate γ_p for stable proteins is determined by their dilution due to cell division, hence we have $\gamma_p = \frac{\ln 2}{\tau}$, where $\tau \sim 20$ min is the cell generation time (see also appendix A.1).

simplicity we assume zero initial conditions

$$m(t=0) = 0 \tag{2.3}$$

$$p(t=0) = 0. \tag{2.4}$$

Both the transcription and translation rate constants k_m and k_p depend on the molecular details of the DNA and mRNA, which influence the amount of produced mRNA and protein for each gene individually. In table 2.1 we gave an upper bound for these two parameters, based on the average elongation rates (1). The system (2.1) – (2.2) of ordinary differential equations (ODEs) together with the initial conditions (2.3) – (2.4) can be solved analytically, yielding

$$m(t) = \frac{k_m}{\gamma_m}\left(1 - \exp(-\gamma_m t)\right) \tag{2.5}$$

$$p(t) = \frac{k_m k_p}{\gamma_m \gamma_p}\left(1 + \frac{\gamma_m \gamma_p}{\gamma_m - \gamma_p}\left[\frac{\exp(-\gamma_m t)}{\gamma_m} - \frac{\exp(-\gamma_p t)}{\gamma_p}\right]\right). \tag{2.6}$$

Steady state of mRNA and protein concentrations is given by the balance between production and degradation rate constants, k_m/γ_p and $k_m k_p/(\gamma_m \gamma_p)$, respectively. In contrast, response times $t_{1/2}$ only depend on degradation rate constants. For mRNA we have $t_{1/2} = \ln(2)/\gamma_m$. If we take into account the typical time scale separation $\gamma_m \gg \gamma_p$ (see table 2.1), we get for the protein response time $t_{1/2} \approx \ln(2)/\gamma_p$. Without time scale separation we can approximate the response time by the sum $t_{1/2} \approx \ln(2)/\gamma_m + \ln(2)/\gamma_p$. In addition to active degradation, proteins and RNAs are diluted by cell division, however this can be taken into account by an effective degradation rate (see appendix A.1).

This model is a coarse grained description, lumping together a complex reaction process into two very simple steps modeled by two linear ODEs. However, it is known that genes are expressed in bursts most probably due to fluctuating promoter activity (46, 18, 167). This implies a noisy gene expression and therefore cells have to deal with the impact varying mRNA and protein levels. Nonetheless, this model is still very useful as a framework to

understand and discuss the process of gene expression. It gives the temporal evolution of average concentrations and will serve as a starting point for a more refined model.

In the following we will outline the molecular details of gene expression. The specific features of DNA and RNA will be described, as well as the nature of the genetic code. The transfer RNAs, which function as a link between the nucleotide sequence of the mRNA and the amino acid sequence of the corresponding protein, are characterized subsequently. We will look more closely at the process of transcription and translation and discuss the organization of bacterial genes in operons. Taking these details into account, we will develop a refined mathematical model of translation. The chapter will be concluded by a discussion of gene expression noise.

2.2. Molecular details of gene expression

2.2.1. DNA and RNA: Information storage and messenger molecules

Primary structure of polynucleotides Nucleic acids, i.e. DNA and RNA, play a predominant role in gene expression (87, 113, 137). Both DNA and RNA are non-branched polymers consisting of nucleotides. These building blocks comprise an organic base, purine or pyrimidine, a 5 carbon sugar, and a phosphate group. Phosphodiester bonds link these nucleotides together, always connecting the $3'$ carbon atom in one sugar with the $5'$ carbon atom in the pentose of the adjacent nucleotide. Therefore all nucleic acids consist of a backbone of repeating sugar-phosphate units, with bases extending as side groups. The synthesis of polynucleotides proceeds only in the $5' \rightarrow 3'$ direction. The directionality and the specific sequence of the four different bases is used by cells to encode the genetic information. The sequence of bases is usually referred to as the primary structure of a DNA or RNA (87).

DNA and RNA differ from each other in three ways (144, 87, 113). The 5 carbon sugar is ribose in case of RNA and deoxyribose in case of DNA. This makes mRNA degrade faster, whereas DNA is chemically much more stable, reflecting their function as messenger and long term information storage molecules, respectively.

The bases adenine (A), cytosine (C) and guanine (G) are common for RNA and DNA, whereas thymine (T) in DNA is substituted by uracil (U) in RNA. These bases can pair with each other by forming hydrogen bonds. Thereby A always pairs with T or U, forming two hydrogen bonds, whereas G and C bind to each other via three hydrogen bonds. These are the canonical or Watson-Crick base pairs. In addition, G and U form base pairs in RNA, whereas theoretically possible base pairs of T with G or C are not found in native DNA (87, 143).

Secondary structure Finally, the nucleic acids differ strongly in their structural properties (144, 113, 87). DNA is commonly found in the form of a stable double helix of two antiparallel DNA strands, held together by hydrogen bonds between complementary bases.

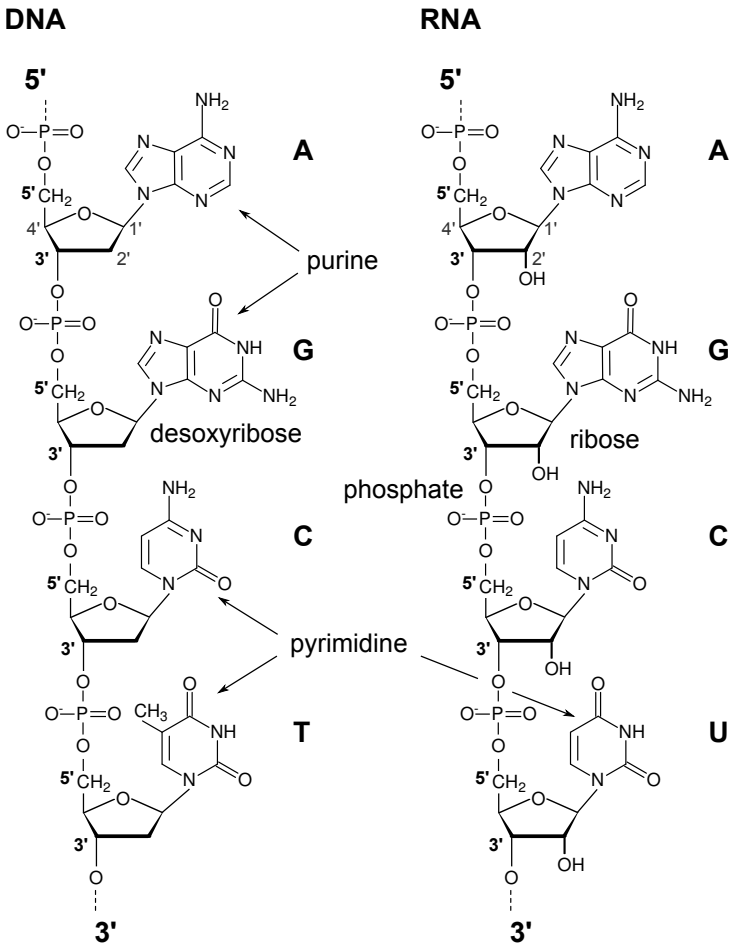


Figure 2.2.: Primary structure of DNA and RNA. Figure adapted from (113).

Hydrophobic and van der Waals interactions of stacked base pairs further stabilize the structure. In contrast, RNA usually is a single-stranded polynucleotide, but can form a huge variety of structures by folding onto itself. At least three levels of organization in RNA structure formation can be distinguished: (1) the primary structure is the specific sequence of bases in a RNA polymer, (2) secondary structure is characterized by the formation of base-pairs between complementary sequences, and (3) the three-dimensional arrangement called tertiary structure (95). Secondary structure formation is usually faster and relies on stronger contacts than tertiary structure (164, 108, 96). Hence RNA folding often can be separated into two steps: first the formation of secondary and then the buildup of tertiary structure. The most common method to predict secondary structure is achieved by finding the structure with minimal free energy. Consider a RNA which folds from the random coil state RC into a structure S_1 (95),

$$\mathrm{RC} \rightleftharpoons \mathrm{S}_1 \ .$$

At equilibrium, the ratio between folded and unstructured conformation is governed by the equilibrium constant K_1

$$K_1 = \frac{S_1}{RC}, \tag{2.7}$$

where S_1 and RC denotes the concentration of the folded and the random coil structure, respectively. Hence a large value of K_1 corresponds to a very stable structure and vice versa. The change in free energy $\Delta G_1 < 0$ due to the formation of S_1 is related to the equilibrium constant K_1 by

$$K_1 = \exp\left(-\frac{\Delta G_1}{N_\mathrm{A} k_\mathrm{B} T}\right), \tag{2.8}$$

where N_A is the Avogadro constant, k_B the Boltzmann constant, and T the absolute temperature. Hence the free energy of a structure is a measure for its stability. If we now consider an alternative structure S_2 with $\Delta G_2 > \Delta G_1$, the difference of free energies quantifies relative contribution of $S1$ and S_2 to the ensemble of structures

$$\frac{[S_1]}{[S_2]} = \frac{K_1}{K_2} = \exp\left(\frac{\Delta G_2 - \Delta G_1}{N_\mathrm{A} k_\mathrm{B} T}\right) > 1. \tag{2.9}$$

At equilibrium, the structure with minimal free energy is therefore the most abundant (95).

Although secondary structure depends on the formation of base-pairs, the overall change in free energy is not so much due to hydrogen bonds but rather due to the stacking of neighboring base-pairs. The latter originates from dipole-dipole induced interactions between the aromatic ring systems of the bases (144). This leads to the formation of helix structures, also referred to as stems. Since a RNA is usually not completely self-complementary, the helix structures are interrupted by regions of unpaired bases, referred to as loops. A whole nomenclature has been developed to describe these structures (fig. 2.3): There are hairpin loops, which close a helix, bulge-loops formed by unpaired bases in one strand in an otherwise double stranded region, internal loops which interrupt a helix by unpaired bases in both strands, and multibranch loops, which connect more than two helices (144). The formation of these loops is penalized by the loss of entropy and therefore energetically unfavorable.

Prediction of secondary structures based on free energy minimization incorporates contributions to free energy changes by an empirical nearest-neighbor model (165, 144, 97, 95). Base-pair stacking energies therefore only take into account the adjacent pairs. Free energy contribution from the hairpin loops depends on the size of the loop and the closing base-pair. In most cases this contribution is only entropic, but there are sequence motifs which are more stable than others, so called tetra-loops, with a nonzero enthalpy. The change in free energy due to bulges and internal loops is mainly determined by their size and the closing base pair. For multibranch loops a linear model is used, taking into account the unpaired bases and the number of helices. In addition to these rules, there are also parameters for dangling ends, closing base-pairs and terminal mismatches. The inset in figure 2.3 shows the calculation of free energy of a stem-loop based on the nearest-neighbor model. Parameter values are

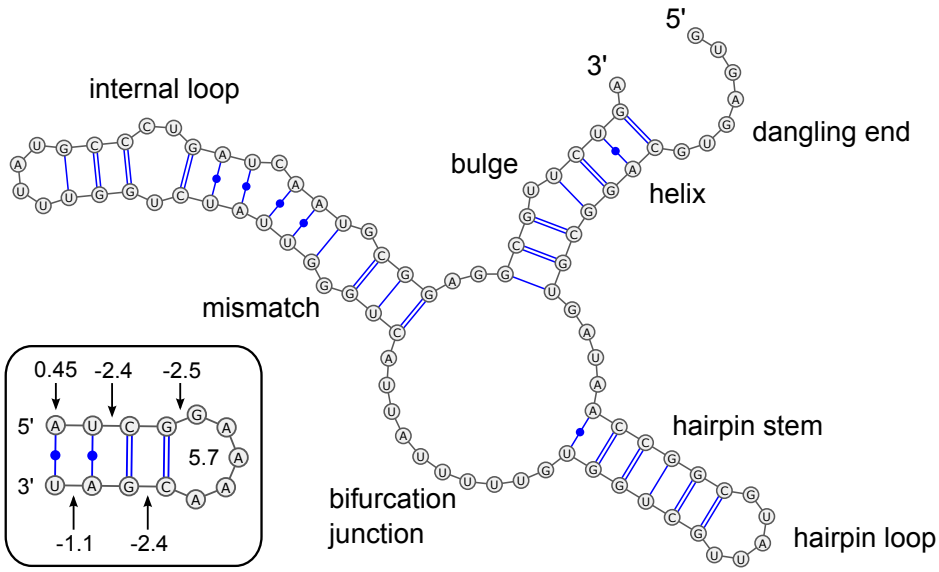


Figure 2.3.: The predicted secondary structure of the *kdpF* RNA. Distinct structural elements are annotated, most importantly distinguishing between regions of paired and unpaired bases. Inset shows how the free energy is calculated by summing up contributions from secondary structure elements like stacking base-pairs or loops. All values are in kcal/mol, giving a total change in free energy of $\Delta G = -2.25$ kcal/mol. The structure was predicted using the ViennaRNA web server and depicted using the VARNA 3.8 software package (59, 28).

taken from the Nearest Neighbor Database hosted by the Turner group at the University of Rochester (156). Closing AU base-pairs are penalized by a positive contribution amounting to +0.45 kcal/mol, the stacking base pairs

$$\begin{array}{ccccc} \text{A} & \text{U} & \text{U} & \text{C} & \text{C} & \text{G} \\ \text{U} & \text{A} & \text{A} & \text{G} & \text{G} & \text{C} \end{array}$$

contribute with $-1.1, -2.4$ and -2.4 kcal/mol. The hairpin loop constrains the conformational space and therefore gives a positive contribution of +5.7 kcal/mol to the change in free energy. However, the loop is stabilized by the closing base-pair GC and the first mismatch formed by G and A leading to a gain in free energy of -2.5 kcal/mol. Hence there is a total change in free energy by $\Delta G = -2.25$ kcal/mol. ΔG enters the expression (2.8) in the exponent which relates the change in free energy and the equilibrium constant. Thus, small differences of the change in free energy have a large impact on the equilibrium.

Tertiary structures form by arranging secondary structure in three-dimensional space, giving rise to triple-base-pairs and pseudoknots. Pseudoknots are formed by pairing bases in the loop of a stem-loop structure and bases not belonging to this structure (see appendix A.2 for a more formal definition). Secondary structures with minimal free energy can be efficiently predicted by excluding pseudoknots, forcing bases of a RNA stretch to only form pairs with

other bases within this sequence. This allows using a recursive scheme, because the minimal free energy of a fragment can be calculated by recursion onto the minimal free energies of smaller fragments. Dynamic programming algorithms use this property by breaking down the determination of the structure with minimal free energy into two steps. In the first step, called recursive fill, the minimal free energy of all fragments is determined, finally yielding the lowest free energy of the whole RNA sequence. Since in this step structures were not generated but only implicitly checked, a second step, referred to as traceback, is necessary to determine the structure with minimal free energy (107, 37, 97) (see also appendix A.2).

RNA molecules not only function as messengers but carry out a wide range of catalytic and regulatory functions. The most prominent example are ribosomes, comprising a complex of several RNAs, referred to as ribosomal RNAs (rRNAs), and proteins. These rRNAs are important for proper recognition of mRNAs and their translation. Another examples of a ribozyme (128, 33) is ribonuclease P which cleaves RNA (51). In addition to the ability to function as enzymes, RNA molecules also can react upon environmental signals, like temperature changes or the presence of small molecules, by altering their structure and consequently controlling translation (75, 128). Recently it has become clear that small RNA molecules play also an important role in specific gene regulation in prokaryotes and eukaryotes (53, 149).

Function and structure of such RNA molecules are closely related, highlighting the relevance of structure prediction. Great advances have been made using the paradigm of free energy minimization for secondary structure prediction. However, as the example of riboswitches shows, RNA molecules may exist in more than one structure. In general we will find a distribution of structures, each with probability

$$p_j = \frac{\exp\left(-\frac{\Delta G_j}{N_\mathrm{A} k_\mathrm{B} T}\right)}{\sum_i \exp\left(-\frac{\Delta G_i}{N_\mathrm{A} k_\mathrm{B} T}\right)}, \tag{2.10}$$

where the structure with minimal energy is the most probable one (95). In addition, folding kinetics may play a role rendering the restriction to the minimal free energy structure insufficient. Furthermore, the formation of pseudoknots may be very important for understanding structure and function of a RNA. The need to overcome these limitations led to the development of new algorithms to predict folding kinetics, suboptimal structures and tertiary structural elements, like pseudoknots (97, 95). However, here we will only focus on the stability of mRNA structures and not on their specific conformations. Since most of the energy resides in the secondary structure, the predicted minimal free energy will be a sufficient proxy to asses the stability of RNA structures (137).

First Base	Second Base: U		Second Base: C		Second Base: A		Second Base: G		Third Base
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Ter	UGA	Ter	A
	UUG*		UCG		UAG		UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG*		CCG		CAG		CGG		G
A	AUU*	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC*		ACC		AAC		AGC		C
	AUA*		ACA		AAA	Lys	AGA	Arg	A
	AUG*	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG*		GCG		GAG		GGG		G

Table 2.2.: The genetic code for bacteria and Archaea according to NCBI (104). It is the same as the standard genetic code, however there are additional start codons indicated by the superscript star $*$.

2.2.2. The genetic code

There are commonly 20 amino acids found in proteins, whereas a mRNA is built up by only four different nucleotides (113, 87, 143). This gives rise to the question how the sequences of amino acids in a protein is encoded in a mRNA. Obviously one and two nucleotides are not enough to encode 20 different amino acids. Hence, at least three nucleotides have to be grouped in order to code for all amino acids. However, triplets of nucleotides permit $4^3 = 64$ different combinations called codons. It turned out that 61 codons, also called sense codons, are specifying amino acids and the remaining three codons, UAG, UAA, and UGA terminate translation, therefore called stop or nonsense codons (106, 143, 113, 87). Thus, the most amino acids, except for methionine and tryptophan, are encoded by more than one codon (table 2.2). The codons corresponding to the same amino acids are referred to as being synonymous. The genetic code is therefore said to be degenerate. Moreover, the code is in general non-overlapping and comma-free. This means that each nucleotide is part of one codon and there are no additional nucleotides between two subsequent codons. Hence there are three ways to group the nucleotides of a given sequence into codons, yielding three different sets of codons and thus completely different amino acid sequences (113). The way

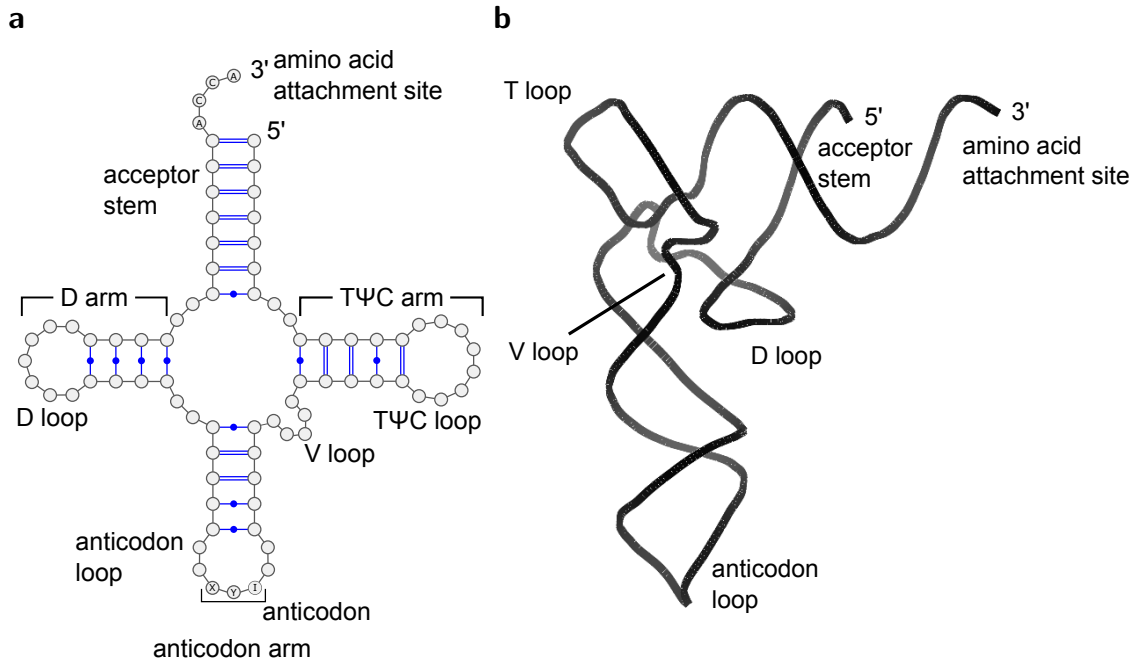


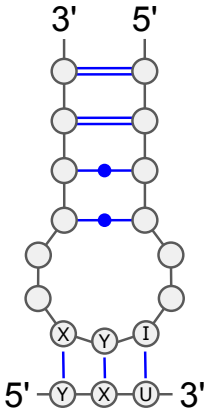
Figure 2.4.: The structure of a tRNA. (a) The cloverleaf like tRNA secondary structure consisting of the acceptor stem, the D arm, the TψC arm and the anticodon arm. The anticodon, which complementary pairs with the codon of the mRNA, resides in the anticodon loop. **(b)** The tertiary structure of the yeast phenylalanine tRNA inferred from X-ray imaging. The helix of the acceptor and the helix of the TψC-arm coaxially stack, the helix of the D-arm and the helix of the anticodon arm coaxially stack (130). The tertiary structure was produced using pymol 1.4.1-1 and X-ray structure data from (132).

of mapping a nucleotide into a codon sequence is called a reading frame, and each mRNA has three possible reading frames. Since usually only one of the reading frames encodes a functional protein, the proper reading frame has to be set during translation initiation. A part of the reading frame containing no stop codons is called an open reading frames (ORF). The other reading frames commonly contain more nonsense codons. Hence ORFs in the other reading frames are usually shorter, thereby terminating translation and preventing the synthesis of large non-functional peptides (87).

It was established by Crick, Brenner and coworkers that the code is triplet, degenerate, non-overlapping, and comma-free (143). Until today only minor exceptions, mostly changing the meaning of non-sense codons, are found, and therefore the genetic code can be assumed to be almost universally valid for all living species on earth (27, 87, 113).

2.2.3. tRNAs effectuate the genetic code

Knowing the mapping from codons to amino acids, we still have to clarify how this is implemented on a molecular level. Francis Crick suggested in 1955 that there are special adapter molecules which carry amino acids and recognize the corresponding codons in the



First Position of Anticodon	C	G	U	I
Third position of Codon	G	C	A	U
		U	G	C
				A
Third position of Codon	G	C	A	U
First Position of Anticodon	C	G	U	I
	U	I	I	G

Figure 2.5.: Codon-anticodon pairing. The tRNA anticodons form complementary base-pairs with the corresponding codon in the mRNA. Note that the anticodon is written in the $3' \rightarrow 5'$ direction, hence the third base of the codon pairs with the base in the first position of the anticodon. Due to wobbling also non-canonical pairs between the third base of the codon and the first base of the anticodon are formed. Table is adapted from (87).

mRNA, thereby linking the sequence of codons to the sequence of amino acids in the protein (26, 58, 143). These adaptor molecules were found to be RNA molecules, therefore called transfer RNAs (tRNAs). They are made up of around 74-95 nucleotides, from which some are complementary to each other, thus forming partial secondary structures by base-pairing (113). Most of the tRNAs are processed after transcription, including removal and addition of nucleotides as well as enzymatic modification of bases. The resulting secondary structure is reminiscent of a cloverleaf comprising four major arms with three of them consisting of stem-loops (fig. 2.4a). However, as revealed by X-ray crystallography (132), tRNAs are not found in this form, but fold into an L like three dimensional structure by forming pseudoknots between internal regions of different arms (fig. 2.4b). The acceptor arm brings together the $5'$ and $3'$ ends of the RNA molecule, thereby forming one end of the L like three dimensional structure of the tRNA. The amino acid gets attached to the $3'$ end which always ends with the nucleotide sequence CCA (87, 113).

Each tRNA carries a specific amino acid added by enzymes called aminoacyl-tRNA synthetases, for each amino acid another one. Due to a proofreading mechanism the error in tRNA charging is only about 1 in $10^4 - 10^5$ (113). The anticodon arm of tRNAs is a stem-loop structure with three nucleotides residing in the loop and forming an anticodon (fig. 2.5). The anticodons of tRNAs form complementary base pairs with the corresponding codon in the mRNA, thereby ensuring the incorporation of the correct amino acid into the polypeptide chain. There are about $30 - 50$ different tRNAs in a cell, thus some amino acids get linked to more than one tRNA. These tRNAs having different anticodons but carrying the same amino acid are called isoacceptors (113). However, there are still more codons than tRNAs, hence some tRNAs must be able to pair with different codons. Most synonymous

codons only differ in their third nucleotide. In addition, it was observed that tRNA species can bind to different synonymous codons (143). These observations led to the postulation of the wobble hypothesis. The anticodon forms canonical Watson-Crick base pairs with the first and second nucleotide of the codon, but may form non-canonical pairs with the third nucleotide. Specifically it was found that adenine is not detected in the first anticodon position, which pairs with the third base in the codon, and appears always to be converted into inosine (I) by post-transcriptional deamination (143). This base can pair with A, C, or T therefore allowing for the flexibility to decode three codons by a single tRNA (87). The table in fig. 2.5 summarizes possible wobble pairs.

2.2.4. Transcription of a gene

Before proteins can be synthesized, the genetic information stored in the DNA has to be transcribed by RNA polymerases into mRNA (fig. 2.6). This process consists of three stages: The initiation, elongation, and termination of the transcription (113, 87).

Bacterial RNA polymerases are huge molecular protein complexes which consist of several subunits. Most importantly, sigma factors associate with the RNA polymerase and control its binding to certain binding sites, so called promoters, in the DNA (115). Which promoters are recognized depends on the specific sigma factor present and thereby leads to the expression of a specific subset of genes appropriate in the environmental context of the cell.

Initiation Once the RNA polymerase has recognized a promoter, it unwinds the double helix and binds to the template strand of the DNA. The correct positioning is accomplished by the interaction with the promoter sequences guiding the RNA polymerase such that its active center is aligned with the beginning of the RNA coding region. Synthesis is initiated by complementary pairing of the first base of the RNA coding sequence in the template strand with a ribonucleoside triphosphate. Further bases are added by complementary pairing to the

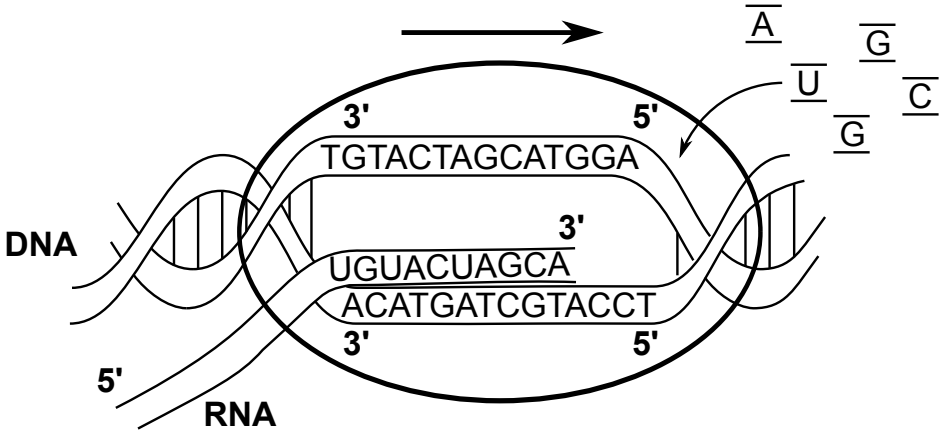


Figure 2.6.: Transcription of DNA into RNA. Figure adapted from (113).

template strand and linkage to the growing polypeptide sequence by phosphodiester bonds in the 5′ to 3′ direction. Thus the synthesized mRNA is complementary and antiparallel to the template strand. After a short transcript (2-6 nucleotides) is synthesized, the RNA polymerase undergoes a conformational change allowing the escape from the promoter region, the release of the sigma factor and the transition to elongation (113).

Elongation As the RNA polymerase elongates along the template strand it unwinds the downstream double helix and correspondingly rewinds the separated DNA strands upstream of its position. This region of unwound DNA, consisting of about 18 nucleotides is called the transcription bubble. Within this bubble, RNA is synthesized step-wise: A Ribonucleoside triphosphate pairs with the complementary base in the single stranded template DNA and is joined to the growing transcript upon cleavage of a diphosphate. The high fidelity of this process is guaranteed by a proofreading mechanism. If an incorrect nucleotide was incorporated, the RNA polymerases moves backwards and removes the last two nucleotides from the transcript (113).

Termination Synthesis ceases when a terminator, which may depend on a specific protein called rho, is transcribed. Rho-dependent terminators rely on the helicase activity of the rho protein: After binding to the 3′ end of the RNA it unwinds the DNA-RNA hybrid and thus stops transcription. In contrast, rho-independent terminators consist of inverted repeats, which after transcription form a hairpin. In addition, the second repeat is followed by a stretch of uracils. It is therefore assumed that hairpin formation and weaker binding of the transcript to the adenine sequence in the template strand facilitates termination (113).

2.2.5. Translation of a gene

Having established the nature of the genetic code, we have to specify how the ribosome binds to the mRNA and sets the correct reading frame, how the corresponding sequences of codons is translated into an amino acid sequence, and how the synthesis of the protein gets terminated. Like in transcription, there are three different steps of translation taking care of this: Initiation, elongation and termination (113, 87, 93) (fig. 2.7). Bacteria are characterized by the lack of cellular compartments. Thus translation can directly start at 5′ end of a mRNA, whose synthesis is still going on. This may lead to a coupling of transcription and translation in bacterial gene expression (143).

Initiation During initiation all components of the translational machinery have to be assembled for allowing protein synthesis: The mRNA carrying the genetic information, the small 30S and the large 50S subunit of the ribosome, three proteins functioning as initiation factors, the initiator tRNA with N-formylmethionine (f-Met) attached to it, and guanosine triphosphate (GTP). The initial step is the binding of the 30S ribosome subunit to the

mRNA. Since only the disassembled ribosome can initiate translation, binding of the initiation factor 3 (IF-3) to the small subunit prevents premature assembly of the ribosome. In addition, initiation factor 1 (IF-1) stimulates dissociation of the large and small subunit. The site to which the ribosomes binds during initiation, the ribosome binding site (RBS), is defined as the region covered by the initiating ribosome and contains about 30 to 40 nucleotides. It comprises the start codon (most common AUG) and a specific sequence motif, named Shine-Dalgarno (SD) sequence. The SD-sequence is complementary to a sequence of nucleotides at the $3'$ end of the 16S rRNA, which is part of the small ribosome subunit (134). By complementary base pairing, this motif facilitates the correct positioning of the small subunit of the ribosome on the mRNA with respect to the start codon. Upon binding of the 30S subunit, the initiator tRNA forming a complex with GTP-activated IF-2 binds to the start codon. Once this complex is assembled, GTP is hydrolyzed and the three initiation factors dissociate from the complex. This allows finally for binding of the large ribosomal subunit, which completes initiation (113).

Elongation After the first tRNA is attached to the start codon residing in the fully assembled ribosome, this 70S initiation complex enters the phase of elongation. In addition to the complex, this requires tRNAs charged with the corresponding amino acids, elongation factors, and GTP. Three binding sites for tRNAs in the ribosome can be distinguished: The E (exit) site, the P (peptidyl) site, and the A (aminoacyl) site (fig. 2.7). After initiation, the initiator tRNA resides at the P site, the only site this kind of tRNA can bind to. The initiation complex then enters the elongation cycle by binding of an aminoacyl-tRNA accompanied by the GTP-bound elongation factor EF-Tu. Once the specific tRNA, whose anticodon complementary pairs with the codon of the mRNA located at the A site, is selected and bound, GTP is cleaved to GDP and the elongation factor bound to GDP is released into the cytoplasm. The growing peptide chain is then bound to the amino acid attached to the tRNA residing in the A site The P site tRNA is vacant and leaves the ribosome through the E site upon which it can be reloaded with the cognate amino acid for a new elongation cycle. Finally, the ribosome moves along the mRNA in $5' \rightarrow 3'$ direction and gets positioned over the next codon. This movement is called translocation and requires the binding of an elongation factor G (EF-G) and hydrolysis of GTP to GDP. Since the attached tRNAs do not move but stay paired to their cognate codons, the tRNA in the A site moves to the P site. This is where the elongation cycle starts to repeat itself: The A site of the ribosome is empty and prepared to bind a new tRNA that is specified by the next codon. The hydrolysis of several GTPs makes the whole process irreversible thus ensuring that the ribosome moves only in one direction along the mRNA (113).

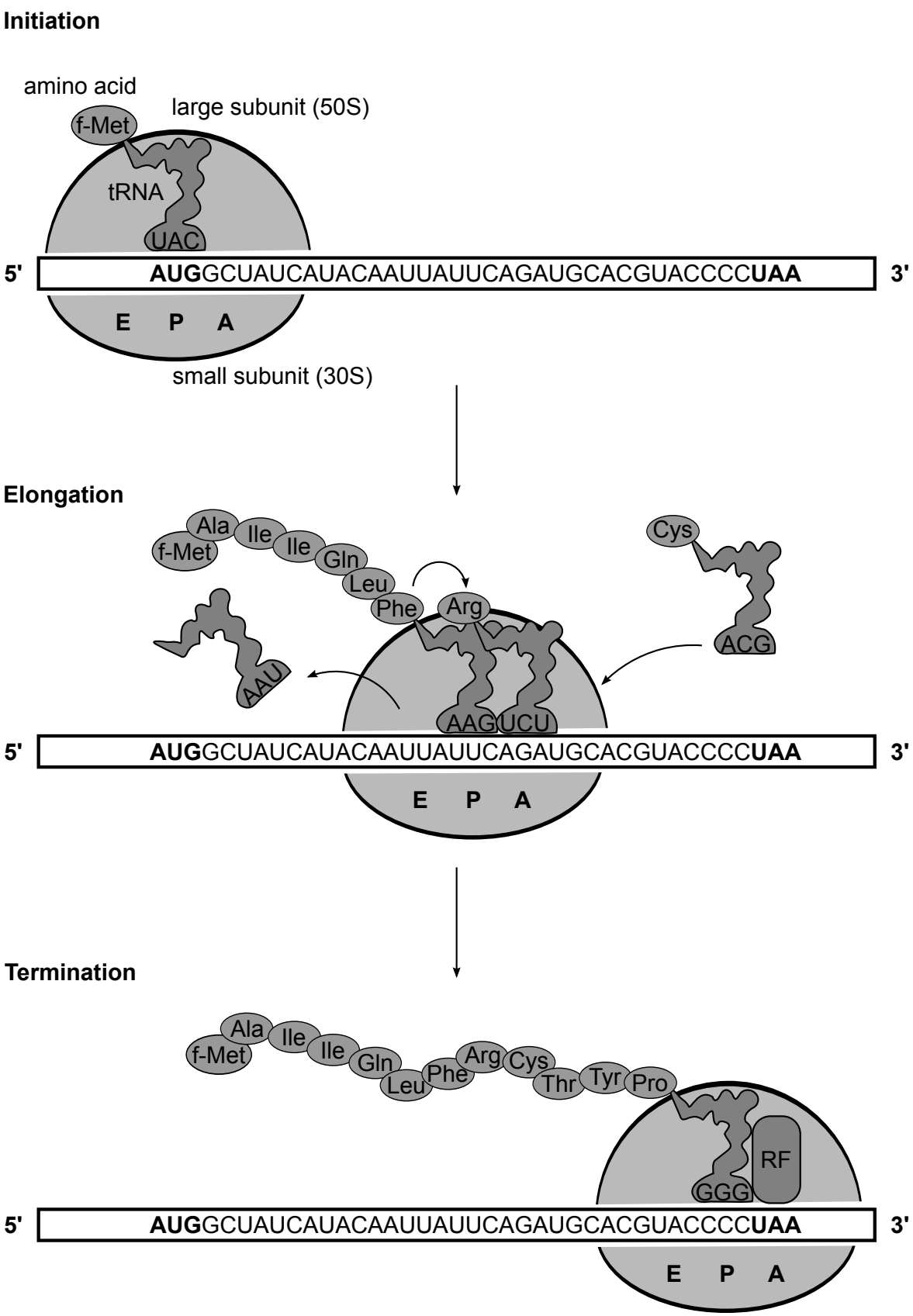


Figure 2.7.: Translation of a gene. Figure adapted from (113).

Termination The whole protein synthesis comes to an end once one of the three termination codons, for which no cognate tRNAs exist, enters the A site of the ribosome. Instead of tRNAs, specific proteins, called release factors (RF-1 and RF-2), bind to the A site (fig. 2.7). This stimulates the cleavage of the polypeptide chain from the tRNA in the P site and its release into the cytoplasm. Upon further hydrolysis of GTP attached to ribosome bound elongation factor G, the elongation complex is disassembled into mRNA, tRNA and ribosomal subunits. These disassembled ribosomes can then initiate translation of the same or another ORF (113, 93).

2.2.6. Organization of a mRNA

In general the transcribed mRNA consist of three distinct regions: A 5′ untranslated region (5′ UTR), the coding sequences and a 3′ untranslated region (3′ UTR) (113). The 5′ UTR can greatly vary in length, and might even be absent (102). Usually it contains the Shine-Dalgarno (SD) sequence around 5-10 nucleotides upstream of the start codon, which is important for the recruitment of ribosomes to the mRNA (134). The 5′ UTR is followed by the codons determining the amino acid sequence of the protein, found in the coding region between the start codon, most likely AUG, and the stop codon. Finally, there is the 3′ UTR of the mRNA, which influences its stability (113).

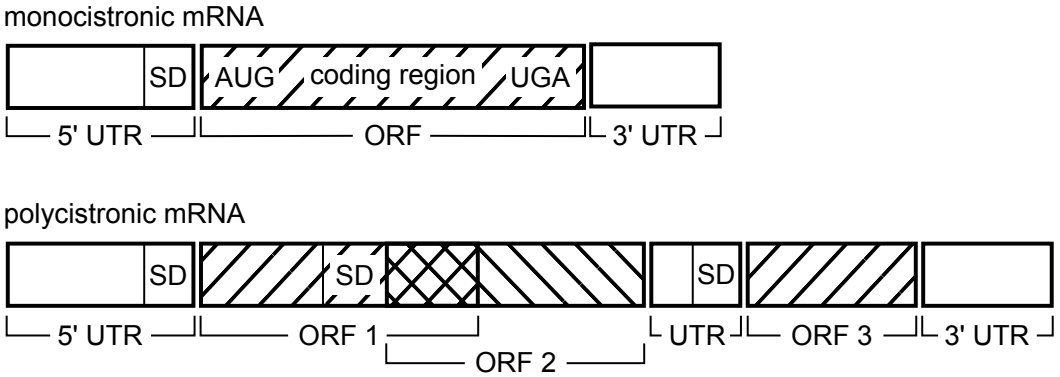


Figure 2.8.: Organization of mono- and polycistronic mRNAs. The monocistronic mRNA usually comprises an untranslated region (UTR) at its 5′ and 3′ end. Upstream of the start codon (here AUG) the so called Shine Dalgarno (SD) sequence is present, which is important for ribosome recruitment to the transcript. The open reading frame (ORF) extends from the start codon to the stop codon, here AUG and UGA. The organization of a polycistronic mRNA is similar. In addition there may exist untranslated regions between two open reading frames, which can also overlap.

However, this simple model for the organization of a mRNA is modified in bacterial cells (113). It is common that genes encoding proteins forming a molecular complex or functioning together in a biochemical pathway are transcribed as one single mRNA. Such a group of genes is called an operon and the corresponding mRNA is termed polycistronic, in contrast to monocistronic mRNAs which only carry the information of a single gene (120, 113).

Polycistronic mRNAs are synthesized when a group of genes is followed by a single terminator, instead of each gene having its own terminator. In addition to the $5'$ and $3'$ UTRs, there might be also untranslated regions between the single genes. However, some genes even overlap in their coding regions. One of the most common motifs found in *E. coli* combines the start codon AUG and stop codon UGA in the sequence AUGA (123).

2.3. Refined model of translation

As we have seen, ribosomes assemble on the transcript during translation initiation of a mRNA. If initiation is slow compared to elongation, we do not have to take into account jamming at the ribosome binding site or along the mRNA. However, if ribosomes are densely packed on the transcript, volume exclusion effects have to be taken into account. In this regime the model of the total asymmetric exclusion process (TASEP) is the appropriate approach to investigate translation. TASEP has been extensively studied as a model system for non-equilibrium statistical physics (101, 172). Moreover, low copy numbers of genes and mRNAs can lead to a noisy expression of gene as we will see in the next section.

Here, we will consider the regime of low ribosome densities and refine the model introduced at the very beginning, which approximates the time evolution of average concentrations within a population of cells by neglecting fluctuations. Thus the rate of translation initiation on the transcripts of the ith gene is given by

$$k_i m_i R, \tag{2.11}$$

where k_i is the rate constant and m_i and R the concentration of the ribosome binding site (RBS) and the free ribosomes, respectively.

After initiation ribosomes move along the mRNA, translating each codon one by one into the corresponding amino acid, thereby elongating the growing polypeptide chain. Each of these steps has a characteristic timescale τ. The elongation rate $v_{ij} = \tau_{ij}^{-1}$ of codon j in the transcript of gene i is predominantly determined by the specificity of the codon-anticodon interactions and abundance of charged cognate tRNAs (170). The latter can be approximated by either the relative abundance of total tRNA levels or tRNA gene copy numbers (154). In the absence of jamming, the total elongation time Δt_i of an open reading frame (ORF) consisting of L_i codons thus can be approximated according to Bulmer (15) by

$$\Delta t_i = \sum_{j=1}^{L_i} \tau_{ij} = \sum_{j=1}^{L_i} v_{ij}^{-1}. \tag{2.12}$$

Hence we get for the overall effective elongation rate v_i

$$v_i = \frac{1}{\Delta t_i} = \frac{1}{\sum_{j=1}^{L_i} v_{ij}^{-1}}. \tag{2.13}$$

A refined scheme for the translation process now looks like

$$\mathrm{m_i} + \mathrm{R} \xrightarrow{k_i} [\mathrm{R_{el}}|\mathrm{m_i}] + \mathrm{m_i} \xrightarrow{v_i} \mathrm{p_i} + \mathrm{R}. \tag{2.14}$$

This is an approximation, since elongation is a multi-step process. The approximation should become better, if the sum in eq. (2.13) is dominated by a single rate constant, i.e. when a rate limiting step exists. In the appendix we solved a simple model for the elongation with discrete steps and compared it to our approximation (see appendix A.3). The concentration $[R_{\mathrm{el}}|m_i]$ of ribosomes attached to transcripts of gene i is governed by the simple equation

$$\frac{d}{dt}[R_{\mathrm{el}}|m_i] = k_i m_i R - v_i [R_{\mathrm{el}}|m_i], \tag{2.15}$$

and the ribosomes have to obey the conservation relation

$$R^T = R + \sum_k [R_{\mathrm{el}}|m_k], \tag{2.16}$$

where R^T is the total concentration. The refined differential equation governing the concentration of proteins therefore now reads

$$\frac{d}{dt} p_i = v_i [R_{\mathrm{el}}|m_i] - \gamma_p p_i, \tag{2.17}$$

which together with eqs. (2.1) and (2.15) comprise our model. However, due to the time scale separation $v_i \gg \gamma_m \gg \gamma_p$, we can apply a quasi-steady state approximation

$$k_i m_i R - v_i [R_{\mathrm{el}}|m_i] \approx 0 \tag{2.18}$$

and thus

$$\frac{d}{dt} p_i = k_i m_i R - \gamma_p p_i. \tag{2.19}$$

The quasi-steady state approximation therefore allows us to determine the protein synthesis rate, whose rate constant k_p is given by

$$k_p = k_i R. \tag{2.20}$$

It is thus the translation initiation rate which governs the synthesis of proteins and consequently the steady state concentration

$$p_i = \frac{k_i}{\gamma_p} m_i R \tag{2.21}$$

in the regime of low ribosome occupancy of an ORF. There is no direct dependence on the elongation rate v_i, but an indirect one through the conservation relation (2.16) for ribo-

somes (15). Especially highly expressed genes might therefore have pronounced effects on the overall translation capacity of a cell. If such genes are slowly translated they sequester ribosomes thereby depleting the pool of free ribosomes. Due to eq. (2.21) this not only affects their own expression but also those of other, even minor translated genes. The specific codon usage of highly translated genes might be therefore a consequence to avoid excessive sequestration of ribosomes which could have a negative effect on cellular fitness (76).

Kinetic model of translation initiation We now want to refine the model with respect to translation initiation. The binding and unbinding of ribosomes to the RBS will be taken into account, thus we consider the more detailed reaction scheme

$$\mathrm{m_i^f} + \mathrm{R} \underset{k_i^-}{\overset{k_i^+}{\rightleftharpoons}} [\mathrm{R_{in}}|\mathrm{m_i}] \xrightarrow{k_i^{cat}} [\mathrm{R_{el}}|\mathrm{m_i}] + \mathrm{m_i^f} \xrightarrow{v_i} \mathrm{p_i} + \mathrm{R} + \mathrm{m_i^f}\ . \tag{2.22}$$

We introduced a new state $[R_{\mathrm{in}}|m_i]$, which denotes the initiating ribosome, bound to the RBS site of ORF i. Once the ribosome is bound it can dissociate or start to elongate along the mRNA, again freeing the RBS. The governing equations now read

$$\frac{d}{dt}[R_{\mathrm{in}}|m_i] = k_i^+ m_i^f R - (k_i^- + k_i^{cat})[R_{\mathrm{in}}|m_i] \tag{2.23}$$

$$\frac{d}{dt}[R_{\mathrm{el}}|m_i] = k_i^{cat}[R_{\mathrm{in}}|m_i] - v_i[R_{\mathrm{el}}|m_i] \tag{2.24}$$

$$m_i = m_i^f + [R_{\mathrm{in}}|m_i], \tag{2.25}$$

where m_i^f denotes the unoccupied RBS. Corresponding to (2.16), a conservation relation for ribosomes holds

$$R^T = R + \sum_j \left([R_{\mathrm{in}}|m_j] + [R_{\mathrm{el}}|m_j]\right). \tag{2.26}$$

We again apply a quasi-steady state approximation, yielding

$$(m_i - [R_{\mathrm{in}}|m_i])R - K_i^M[R_{\mathrm{in}}|m_i] \approx 0 \tag{2.27}$$

$$k_i^{cat}[R_{\mathrm{in}}|m_i] - v_i[R_{\mathrm{el}}|m_i] \approx 0, \tag{2.28}$$

where we used eq. (2.25) and introduced the Michaelis-Menten constant

$$K_i^M = \frac{k_i^- + k_i^{cat}}{k_i^+}. \tag{2.29}$$

The solution reads

$$[R_{\text{in}}|m_i] = \frac{m_i R}{K_i^M + R} \tag{2.30}$$

$$[R_{\text{el}}|m_i] = \frac{k_i^{cat}}{v_i} \frac{m_i R}{K_i^M + R}. \tag{2.31}$$

Therefore the equation governing the protein concentration now takes the form

$$\frac{d}{dt} p_i = k_i^{cat} \frac{m_i R}{K_i^M + R} - \gamma_p p_i. \tag{2.32}$$

In the limit $K_i^M \gg R$ we obtain eq. (2.19), with $k_i = k_i^{cat}/K_i^M$.

Thermodynamic model of translation initiation If $k_i^{cat} \ll k_i^-$ holds, the complex formation is faster than the commitment to elongation and therefore the association-dissociation reaction achieves a rapid equilibrium (9). In this case the introduced Michaelis-Menten K_i^M constant equals the dissociation constant K_i^D

$$K_i^M = \frac{k_i^- + k_i^{cat}}{k_i^+} \overset{k_i^{cat} \ll k_i^-}{\approx} \frac{k_i^-}{k_i^+} = K_i^D. \tag{2.33}$$

In this regime a thermodynamic model for the complex formation can be applied (124). The equilibrium constant k_i^+/k_i^- is related to the Gibbs free energy ΔG_i, released upon binding of the ribosome to the RBS, by

$$\frac{k_i^+}{k_i^-} = \frac{1}{K_i^D} = \exp\left(-\frac{\Delta G_i}{N_{\text{A}} k_{\text{B}} T}\right), \tag{2.34}$$

where N_{A} is the Avogadro constant, k_{B} the Boltzmann constant, and T the absolute temperature (9). Plugging this into eq. (2.30) we obtain

$$[R_{\text{in}}|m_i] = \frac{m_i R \exp\left(-\frac{\Delta G_i}{N_{\text{A}} k_{\text{B}} T}\right)}{1 + R \exp\left(-\frac{\Delta G_i}{N_{\text{A}} k_{\text{B}} T}\right)}. \tag{2.35}$$

For $R \exp\left(-\frac{\Delta G_i}{N_{\text{A}} k_{\text{B}} T}\right) \ll 1$, implying that only a minor fraction of RBS is occupied by ribosomes, this simplifies to

$$[R_{\text{in}}|m_i] = m_i R \exp\left(-\frac{\Delta G_i}{N_{\text{A}} k_{\text{B}} T}\right). \tag{2.36}$$

In this limit the rate equation for proteins is given by

$$\frac{d}{dt} p_i = k_i^{cat} m_i R \exp\left(-\frac{\Delta G_i}{N_{\text{A}} k_{\text{B}} T}\right) - \gamma_p p_i \tag{2.37}$$

and therefore the initiation rate constant equals

$$k_i = k_i^{cat} \exp\left(-\frac{\Delta G_i}{N_\mathrm{A} k_\mathrm{B} T}\right). \tag{2.38}$$

2.4. Gene expression noise

The amazing biological variability predominantly arises because of genetic differences, but also history and environmental fluctuations lead to different phenotypes. More surprising, even genetically identical cells being exposed to the same environment and sharing a common past may show marked variation of their phenotypical characteristics. Random variability in the synthesis of gene products, referred to as gene expression noise, is commonly held liable for such phenotypic variations of clonal cells (29, 117) and can be attributed to intrinsic and extrinsic sources (150, 40, 116, 117, 91, 118).

Intrinsic noise may arise due to fluctuations in biochemical reactions, which are inherently noisy. The stochastic variability only averages out in the limit of large molecule numbers, resulting in deterministic dynamics. However, if only a small number of interacting molecules is present, stochasticity becomes apparent. As the main players, i.e. genes and transcripts, may be present and active in small numbers, the abundance of gene products can be subject to random variations. Consequently, the stochastic events during gene expression, from the activation of the promoter to the degradation of gene products, even differ between two identical copies of a gene in a single cell. Most important, mRNAs are synthesized in stochastic bursts due to the random switching of the gene's promoter activity (46). Since typically several proteins are translated from a single transcript, the mRNA bursts are amplified and result in corresponding protein bursts (18, 167). However, if the protein lifetime is much larger than the time between two consecutive production burst, accumulation of proteins averages out the random variability (38).

In contrast, extrinsic noise arises due to fluctuation of factors which influence gene expression in a single cell on a global scale. These include the overall transcriptional and translational capacity as well as the abundance of gene-specific transcription factors, themselves being subject to intrinsic noise (91). Such fluctuations are propagated downstream and cause cell-to-cell variations in a population.

The contribution of intrinsic and extrinsic noise to the total level of noise was tested experimentally in *E. coli* (fig. 2.9). To this end two fluorescent reporter genes, the yellow fluorescent protein (*yfp*) and cyan fluorescent protein (*cfp*) genes, were placed under the control of identical promoters in the same chromosome (40). A scatter plot of measured fluorescent intensities reveals the relative contribution of extrinsic and intrinsic noise. Uncorrelated fluctuations, originating from intrinsic noise, cause the deviation of points perpendicular to the diagonal line along which CFP and YFP intensities covary. In contrast, correlated fluctuations, resulting from extrinsic noise, leads to the spread of fluorescent intensities along

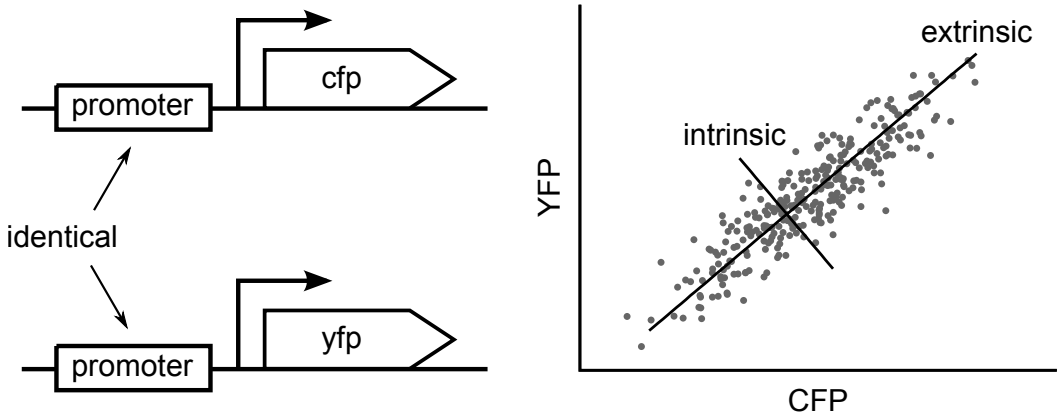


Figure 2.9.: Intrinsic and extrinsic noise. Scheme of the classical experiment by Elowitz et al.: Two different fluorescent protein genes, *yfp* and *cfp* were put under the control of the same promoter. Correlated variations within a population arise due to differences in the global cellular state and in the amount or activity of gene specific regulatory proteins. All stochastic events inherent to the biochemical reactions of the gene expression process contribute to intrinsic noise (40).

the diagonal line.

It is clear that noise perturbs the functioning of cellular pathways, like signaling networks and genetic circuits, by altering expression levels and consequently rates of chemical reactions. Thus evolution developed strategies to control gene expression noise and mitigate its effects. A study using reporters with modulated translation and transcription efficiency demonstrated that noise in protein levels predominantly increases with enhanced translational efficiency, but depends only weakly on the modulation of transcriptional efficiency (110). Thus cells may employ high transcription but low translation rates in order to reduce expression noise of important proteins. Controlling the gene copy number is another method to lower the level of intrinsic noise, which is predicted to scale with the inverse square root of the gene copy number (117). It has been proposed theoretically that cells may use autoregulatory negative feedbacks to lower noise levels. Indeed, design of such a genetic circuit demonstrated its superiority to unregulated systems (10). However, a recent theoretical study stressed the fundamental limits of regulatory feedback mechanism. Regulatory molecules have to be produced at least as frequently as the controlled component to fulfill their function. It turns out that noise is reduced only with the fourth root of the number of signaling events, making it extremely expensive to increase accuracy (78).

A completely different strategy to cope with the presence of noisy expression levels was taken by the evolution of pathway topologies which maintain function despite randomly varying component levels (73, 133, 145). One such example is the chemotaxis pathway of *E. coli*, which can buffer correlated fluctuations of its constituting proteins (73). It is specifically the presence of a phosphatase which makes the adapted output of this signaling network robust against extrinsic noise. Moreover, as we have seen in section 2.2.6, genes are often

cotranscribed as a single polycistronic mRNA, resulting in a coupling on the transcriptional level. Hence the expression of those genes is correlated even if the promoter exhibits stochastic activity. Consequently, components of the chemotaxis pathway are expressed from two operons in *E. coli* (103).

3. Translation initiation and codon usage

An abbreviated version of this chapter is to be published as "Suppression of mRNA structure shapes codon usage at gene start in bacteria". Contributing authors: Kajetan Bentele (Theory and Experiment), Paul Saffert (Experiment), Robert Rauscher (Experiment), Zoya Ignatova, and Nils Blüthgen.

3.1. Introduction

Which evolutionary constraints shape the genome of an organism? This question has been puzzling researchers for many decades (131, 55, 153, 65, 106, 63). Clearly, the protein-coding and promoter sequences, regulating the expression of the former, are the main determinants of genomic sequences. However, even for a given amino acid sequence, messenger RNAs (mRNAs) may differ, since the genetic code is degenerate: Except for tryptophan and methionine, the amino acids are encoded by two, four or six different codons (27, 113). But what are the mechanisms that led to the choice of specific codons?

On a genome-wide scale some codons are preferred to others, termed global codon usage bias (49, 131). While the origin of the bias is not completely clear, it is known that each genome has a specific bias, resulting from a balance between selection, mutation and genetic drift (54, 129). Driving forces for preferentially selecting specific codons could be efficiency and accuracy of translation (45, 19, 129, 35, 162, 154, 153), GC-content (90), environmental factors (135, 90), or DNA folding (56). The species-specific codon bias is reflected by the amount of the cognate transfer-RNAs (tRNAs), i.e. frequently used triplets will demand more tRNAs whereas rare triplets are read mostly by low abundant tRNA species (60, 34). Triplets read by major tRNAs are therefore most likely translated with higher speed than the codons read by minor tRNAs (170). Highly expressed genes thus display a stronger bias towards usage of highly abundant codons reflecting tRNA availability (131, 155).

Within each gene, rare codons are far from being uniformly distributed along the coding mRNA sequences; they tend to cluster and may transiently retard ribosomal traffic. Between structural domains rare codons slow down ribosomal speed at the domain boundaries allowing accurate folding of the structural domains in proteins (169, 74). Furthermore, rare codons may exert regulatory function under starvation (39, 168), but may also cause misfolding in highly expressed genes and are therefore avoided (162).

Genome-wide analysis of some selected organisms revealed that the codon usage within the

first few codons of a coding sequence differs from the usage elsewhere in the genome (42, 154). These findings suggest a specific evolutionary pressure for the selection of codons at the start of a gene. According to the "ramp hypothesis", rare codons which correspond to low abundant tRNAs may be preferentially used to reduce elongation speed at the beginning of a gene (154). This may be advantageous because it could reduce the likelihood of ribosomal "traffic jams" along the mRNA, which may give rise to protein bursts (32) or premature termination of the translation.

Before ribosomes can start to elongate, however, they have to bind to the mRNA thereby initiating translation. As outlined in the background chapter 2, the translation initiation rate can be expressed in terms of the change in Gibbs free energy ΔG_{init} upon binding of the ribosome (eq. (2.38)), more specifically its 30S subunit, to the transcript (124)

$$k = k^{cat} \exp\left(-\frac{\Delta G_{\text{init}}}{N_{\text{A}} k_{\text{B}} T}\right). \tag{3.1}$$

The change in free energy ΔG depends on the specific sequence around the translation initiation codon and can be broken down into different contributions, with the reference state being the fully unfolded local sequence ($G = 0$)

$$\Delta G_{\text{init}} = \Delta G_{\text{mRNA:rRNA}} + \Delta G_{\text{start}} + \Delta G_{\text{spacing}} - \Delta G_{\text{standby}} - \Delta G_{\text{mRNA}}. \tag{3.2}$$

Here, $\Delta G_{\text{mRNA:rRNA}} < 0$ is the energy released upon hybridization and co-folding of the Shine Dalgarno sequence, a sequence motif located around 8 nt upstream of the start codon, to the $3'$ terminal sequence of the *E. coli* 16S rRNA (134). If the Shine Dalgarno sequence is located too close or too far away from the start codon, the 30S complex becomes distorted, which is taken into account by the penalty term $\Delta G_{\text{spacing}} > 0$. When the initiating tRNA anticodon loop binds to the start codon, the free energy $\Delta G_{\text{start}} < 0$ is released. An additional contribution comes from the folding of mRNA around the start codon. The change in free energy $\Delta G_{\text{mRNA}} < 0$ is the work required to unfold mRNA from its most stable secondary structure state, referred to as the minimum free energy structure. Finally, $\Delta G_{\text{standby}} < 0$ stems from any secondary structure sequestering the standby site, i.e. the four nucleotide upstream of the Shine Dalgarno sequence (124).

The impact of mRNA folding (ΔG_{mRNA}) on the translation initiation rate and consequently the global rate of translation has been shown previously (98). Increasing the stability of mRNA folding in the region containing the ribosome binding had dramatic effects: A single nucleotide substitution led to the decrease of expression by a factor of 500 in expression (30). Recently, a synthetic library of 154 genes that varied randomly at synonymous sites, but all encoded the same green fluorescent protein (GFP), was engineered (76). The variation of expression levels could be to a large extent explained by stability of mRNA folding near the ribosome binding site. In contrast, the selective choice of synonymous codons along the entire transcript (codon bias) did not correlate with gene expression. Similarly, a small

but significant anti-correlation between ribosome density throughout the transcript and the amount of mRNA secondary structure 10 nucleotides upstream of translation start site was found in the yeast transcriptome (70). Moreover, the reduction of secondary structure at the translation initiation site is a shared feature in many genomes (50).

The folding energy ΔG_{mRNA}, which will be denoted by G for brevity, is directly influenced by the nucleotide sequences downstream of the translation start codon. This raises the intriguing question as to whether the specific codon bias at the beginning of genes may have evolved to reduce the folding energy of the mRNA sequence around the translation start site (42, 114). While according to the "ramp hypothesis" the choice of codons is shaped by the need to slow down early elongation, our "structure hypothesis" emphasizes the importance of translation initiation. We investigated the fundamental question of differential codon usage at gene start by systematically analyzing 414 bacterial genomes. We found that codon usage deviates only in the first few codons if the genome is GC-rich and thus folding energy of mRNA is large. Only rare codons which reduce GC-content are preferentially selected at gene start, suggesting that codon usage at the beginning of genes is driven by the pressure to reduce secondary structure at translation initiation sites. The hypothesis is further corroborated by evolutionary simulations and experimental measurements on the translational efficiency of two *E. coli* genes with various synonymous starting sequences, underpinning the functional relevance. In addition we investigate theoretically the impact of slowly translated codons in the early phase of elongation on protein synthesis by employing the simple model developed in chapter 2.

3.2. Results

3.2.1. Unusual codon usage around the translation start site in bacteria

The sequences from *E. coli* were collected from the database EcoCyc (71), and all protein coding genes were aligned with respect to the translation start. For all sequences, we calculated the frequency of synonymous codons, i.e. the frequency conditioned on the amino acid, at each position downstream of the start codon, which we term local codon usage. This frequency was then compared to the overall frequency of synonymous codons, i.e. global codon usage. To quantitatively assess the deviation of the local codon usage at each position from the global codon usage, we calculated the Kullback-Leibler divergence (KLD) at each position (fig. 3.1),

$$\mathrm{KLD}(k) = \sum_{i=1}^{20} \sum_{j=1}^{S_i} p_{i,j}(k) \ln \frac{p_{i,j}(k)}{q_{i,j}}. \tag{3.3}$$

Here, $q_{i,j}$ denotes the global and $p_{i,j}(k)$ the local synonymous codon frequency at position k (see appendix B.1 for details). Larger KLD values indicate larger differences between local and global codon usage. Figure 3.1 shows the KLD for *E. coli* as a function of the codon

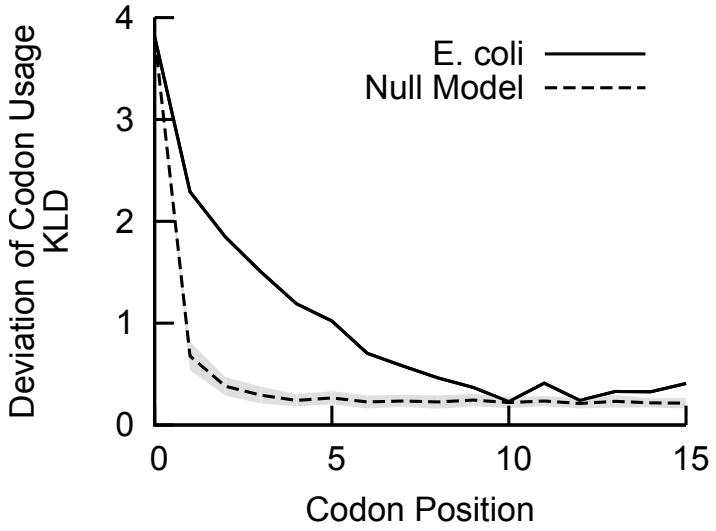


Figure 3.1.: Deviation of codon usage at beginning of genes in *Escherichia coli*. In *E. coli*, the frequency of synonymous codons within the first 8 codons after translation start deviates significantly from the global codon usage in the genome. The deviation was quantified using the Kullback-Leibler divergence (KLD). The bias of the KLD due to finite size sampling was estimated using a null model, which conserved amino acid sequences (SSC). Grey area shows the mean estimated from the null model ± standard deviation.

position (solid line). Due to finite sampling, the KLD is typically larger than zero even if the local and global codon usage do not differ (55, 122). Hence we estimated finite sample effects using a null model based on randomly generated sequences, by shuffling synonymous codons within a gene (null model SSC). Clearly, when compared to the null model, the local codon usage (solid line) deviates strongly from the global codon usage within the first 4-6 codons (fig. 3.1). There is a significant deviation within the first 8 codons and no significant difference thereafter. The currently proposed explanation of this deviation of the codon usage at the beginning of a coding sequence is to prevent jamming of the ribosomes along the mRNA (154), suggesting that an unusual codon usage proximal to the initiation codon would be a universal feature for bacteria. We thus investigated whether different prokaryotes show such strong deviation in codon usage within the first five codons and collected sequence data of 414 bacteria covering 311 species out of 182 different genera from the database BioCyc (66). For all 414 bacterial genomes, we calculated the average KLD for the first five codons after the translation initiation codon. To remove effects from the finite sample size, we subtracted the mean KLD calculated from our null model (fig. 3.2a). The resulting score, ΔCU, provides a quantification of the extent to which local and global codon usage differ within the first five codons. Surprisingly, most bacteria show less pronounced deviation from the codon usage than *E. coli* (fig. 3.2b). For example, *Thermoanaerobacter tengcongensis* (*T. tengcongensis*) lacks almost any bias in the codon usage of the first codons (fig. 3.2b, top inset). *Bacillus subtilis* (*B. subtilis*) shows an intermediate deviation (fig. 3.2b, second inset

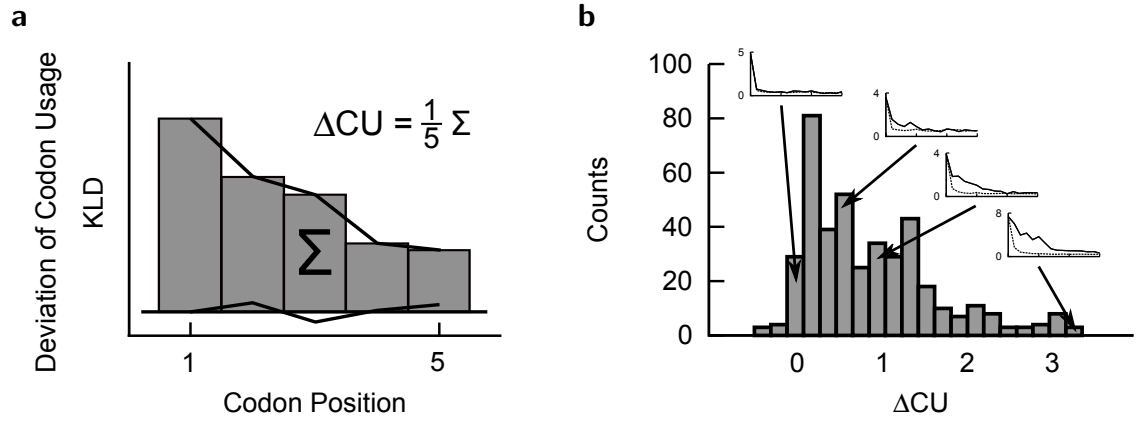


Figure 3.2.: Deviation of codon usage at beginning of genes is widespread among bacteria. (**a**) Deviation of codon usage at the beginning of genes was quantified by the average KLD, after subtracting the bias, within position 1 through 5. (**b**) This deviation, ΔCU, scatters over a range from about 0 to 3. Insets from left to right correspond to KLD profiles calculated for the genomes of *Thermoanaerobacter tengcongensis*, *Bacillus subtilis*, *E. coli* and *Aeromonas hydrophila*, respectively.

from the top). At the other extreme, *Aeromonas hydrophila* (*A. hydrophila*) exhibits very strong deviation (fig. 3.2b, bottom inset) of the local codon usage. These results suggest two alternatives: Either different bacteria exert different pressures to prevent jamming of the ribosomes along the mRNA and therefore show different deviations from the codon usage, or preventing ribosomal jams is not the major driving force for selection of the first codons within genes.

3.2.2. Suppression of secondary structure around translation start site depends on global GC-content

There is a universal trend of reduced mRNA stability around the translation-initiation site in the genomes of prokaryotes and eukaryotes (50). We hypothesized that reduced secondary structure around the translation start might twist codon choice towards unusual codons. We therefore investigated the propensity of mRNA folding around the translation start in bacteria. To quantify the mRNA secondary structure in a position-specific manner, we calculated the folding energy for stretches of 39 nucleotides around each position using the Vienna RNA Package (59) (fig. 3.3). The length of the mRNA-stretch of 39 nucleotides was chosen based on the reported typical length of mRNA the ribosomes bind to (13). The folding energy profiles generated for each mRNA were aligned at the start codon, and a plot showing upper and lower quartile of the energy distribution reveals a strong tendency towards reduced secondary structure around the start codon (fig. 3.3). Away from the gene start, typical folding energy of 39 nucleotide stretches of mRNA is around $G \approx -7.5$ kcal/mol. Interestingly, individual sequences show strong variability, covering the range from -15 to

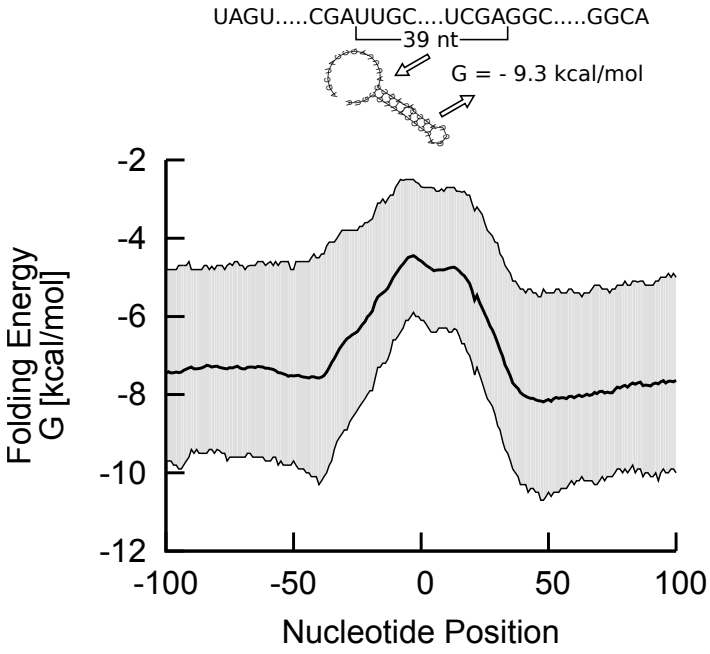


Figure 3.3.: Suppressed mRNA folding around the gene start in *E. coli*. Folding energies of the *E. coli* mRNA sequences calculated within a sliding window of 39 nucleotides. The average folding energy as a function of nucleotide position is shown as the thick black line. The grayed area shows the inter-quartile range. Around the start codon folding energy increases, indicating the suppression of mRNA secondary structure.

0 kcal/mol. However, around the translation start site there is a region of approximately 20 nucleotides where the secondary structure is strongly reduced, with an average folding energy of about $G \approx -5$ kcal/mol. Furthermore, the folding energy around the start codon is more confined, as the smaller spread shows (fig. 3.3).

The energy profiles generated with the 39 nucleotide-sliding windows quantify only short-range secondary structures. Thus, the observed profiles do not necessarily exclude that the ribosome binding region might be involved in strong long-range secondary structures. For all mRNAs of *E. coli*, we next calculated the probability of each nucleotide position to be unpaired. Similar to the energy profile, we observed that the probability for nucleotide to be unbound is significantly increased around the start codon compared to the rest of the coding sequence (fig. B.1a in appendix B.3).

Next, we investigated whether the suppression of secondary structure is a general feature for all bacteria and calculated the folding energy profiles for all mRNAs of 414 bacteria. To compare different bacteria, we calculated the average folding energy G for each bacterium at every position relative to the start codon, and defined the energy at the translation start site, G_0, as the average of G within a window of 11 nucleotides around the gene start (fig. 3.4a). To compare it with the typical folding energies of mRNA in the organism, we defined a baseline folding energy, G_{bl}, as the mean within a 50 nucleotide large window starting from nucleotide position 150 relative to the start codon. The difference ΔG between G_0 and G_{bl}, provides a

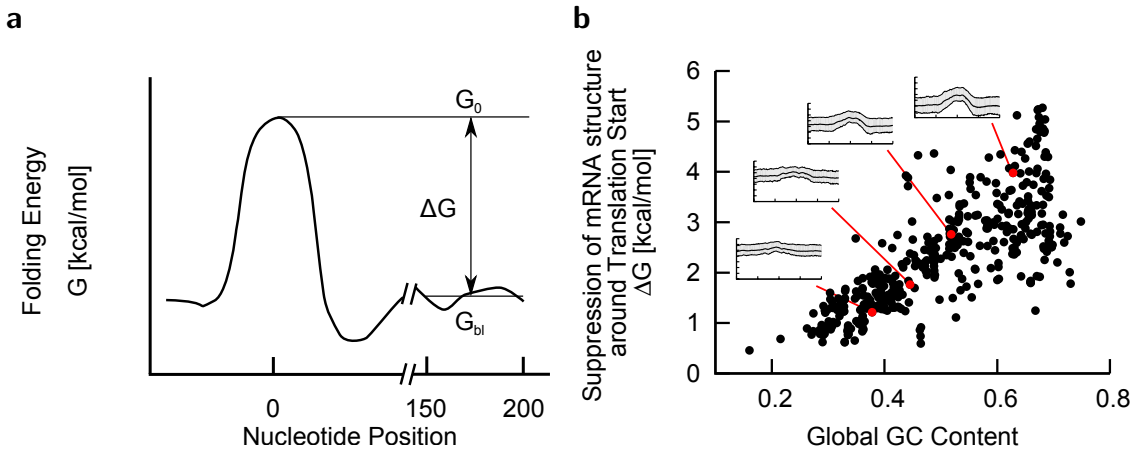


Figure 3.4.: Suppression of mRNA structure depends on global GC-content of genome. (**a**) Schematic representation of the average folding energy. The baseline folding energy G_{bl} was obtained by calculating the mean within a 50 nucleotide large window starting from nucleotide position 150 relative to the start codon. The deviation from this average folding energy around the start codon was characterized by the difference $\Delta G = G_0 - G_{\mathrm{bl}}$. The folding energy at the translation start site, G_0, is obtained by calculating the mean within a window of 11 nucleotides around the start codon. (**b**) The suppression of the folding around the start codon, ΔG, depends on the GC-content. The deviation is stronger the higher the GC-content, i.e. the more stable mRNA folds elsewhere. The insets from bottom to top correspond to the bacteria *Th. tengcongensis*, *B. subtilis*, *E. coli*, and *A. hydrophila*, respectively. Axes of the insets as in fig. 3.3, but y-range goes from -14 to 0 kcal/mol.

quantitative measure for the reduction of the secondary structure (fig. 3.4a). Baseline folding energy of coding regions varies very strongly among bacteria, from approximately -14 to -2 kcal/mol. Not surprisingly, the average folding energy is largely determined by the genomic GC-content (fig. B.1b in appendix B.3).

We interpreted the suppression of secondary structure around the gene start as a result of an evolutionary pressure towards increasing the accessibility of the ribosome binding region. Since the average folding energy varies to a large extent between organisms, we expect that the pressure to reduce the mRNA folding around the start codon will also strongly differ. For organisms in which the high GC-content increase the propensity of formation of strong secondary structures we would expect a specifically strong reduction of folding around the start codon. Indeed, the reduction of secondary structure around the start codon, ΔG, depends on the GC-content in the respective bacterium (fig. 3.4b). Bacteria with a baseline energy around -6 to -2 kcal/mol corresponding to a global GC-content smaller than about 0.45, e.g. *T. tengcongensis* and *B. subtilis*, show hardly any reduction of secondary structure ($\Delta G < 2$ kcal/mol, shown as two insets at bottom of fig. 3.4b). Bacteria with a GC-content of ≈ 0.5 and a baseline energy of $G_{\mathrm{bl}} \approx -8$ kcal/mol, like *E. coli* (fig. 3.4b, second inset from the top), show a moderate decrease of secondary structure with $\Delta G \approx 2 \ldots 3$ kcal/mol.

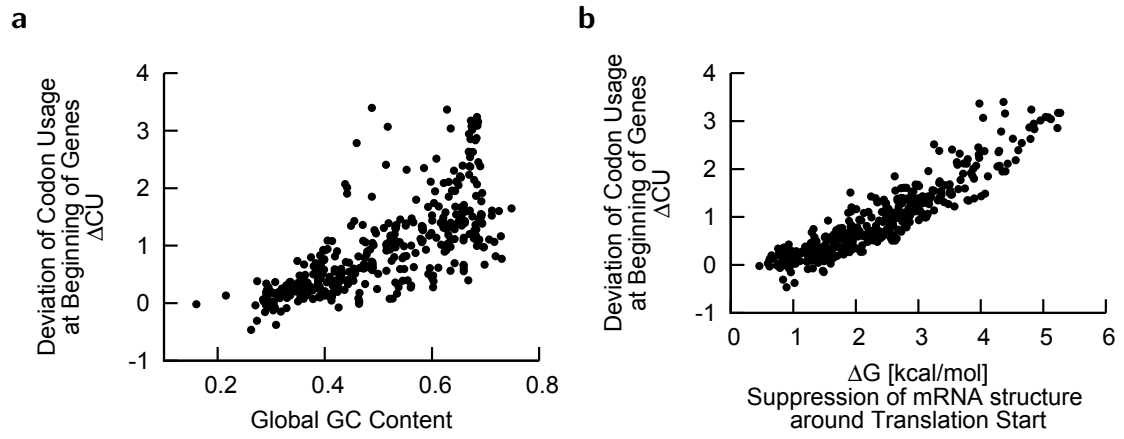


Figure 3.5.: (**a**) Deviation of codon usage at the beginning of genes, ΔCU, correlates with the global GC-content. (**b**) Deviation from the global codon usage, ΔCU, and decreased mRNA structure, ΔG, around the start codon are strongly correlated. The figure shows results for all genes within operons (414 bacterial genomes, correlation coefficient $r = 0.93$)

In contrast, *A. hydrophila* (fig. 3.4b, top inset) with a higher GC-content (≥ 0.6) and thus strong average structures ($G_{\mathrm{bl}} < -10$ kcal/mol) show an increase of folding energy by $\Delta G > 3$ kcal/mol.

3.2.3. Selection of unusual codons correlates with the reduction of secondary structure

The bias towards using unusual triplets proximal to the start codon varies considerably between bacterial species. The secondary structure suppression around the gene start differs strongly, and correlates with the GC-content of an organism (fig. 3.4b). In addition, the deviation of codon usage also correlates with the global GC-content, similar to suppression of structure (fig. 3.5a). We therefore investigated whether the observed unusual codons within the beginning of protein-coding mRNA sequences are selected to decrease the secondary structure propensity around the start codon. Interestingly, we found that ΔCU (unusual codon usage) is strongly correlated with ΔG (suppression of secondary structure) with a correlation coefficient of $r = 0.93$ (fig. 3.5b). The correlation remained when we excluded overlapping genes from our analysis (fig. B.1c in appendix B.3). This result suggests that unusual codon usage may have evolved to suppress the secondary structure at gene start.

Mechanistically, the choice of specific codons could suppress secondary structure by reducing the GC-content locally. Indeed, the GC-content in *E. coli* was strongly reduced within the first 4-6 codons (fig. 3.6a). Synonymous codons differ mostly in their third base and as expected GC3-content, i.e. GC-content at the third nucleotide position, is strongly decreased. In addition, the GC-content in the first and second nucleotide position (GC1- and GC2-content, respectively) is also decreased within the first 4-6 codons (fig. 3.6b). The de-

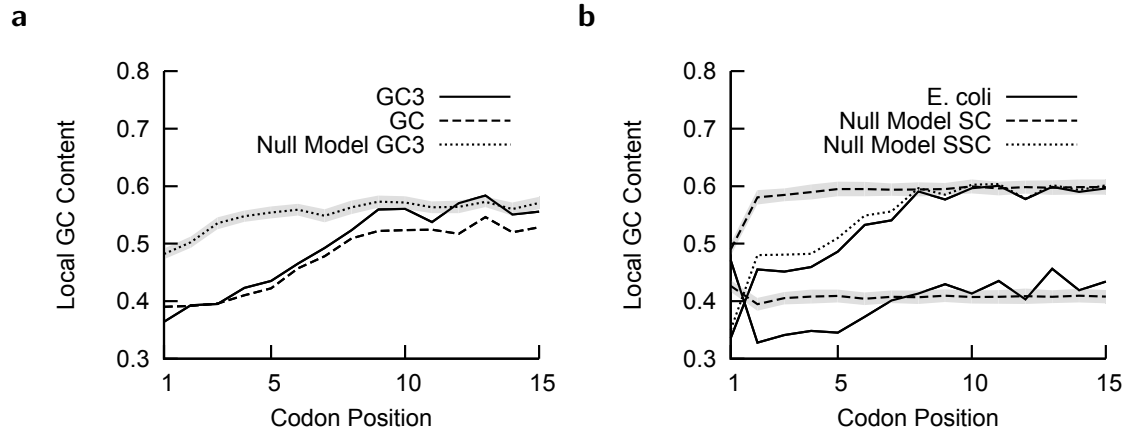


Figure 3.6.: GC-content at the beginning of genes in *E. coli*. (**a**) GC-content and GC-content at third nucleotide position of codons (GC3-content) decrease at the beginning of genes in *E. coli*. Grey area shows mean GC3-content ± standard deviation, estimated from the null model (SSC). (**b**) Also GC1- and GC2-content is decreased at gene start in *E. coli*. Since GC-content for most codons does not differ at first position, there is hardly any difference between the native sequences and the sequences obtained by shuffling synonymous codons (null model SSC). However, GC1- and GC2-content of native sequences significantly differs from what is yield by shuffling codons (SC).

crease in GC1- and GC2-content is also visible with null models with conserved amino-acid sequences (SSC) but almost vanishes if all codons are shuffled (SC), which suggests that amino-acids encoded by AU-rich codons are chosen preferentially at gene start (fig. 3.6b, dotted and dashed line).

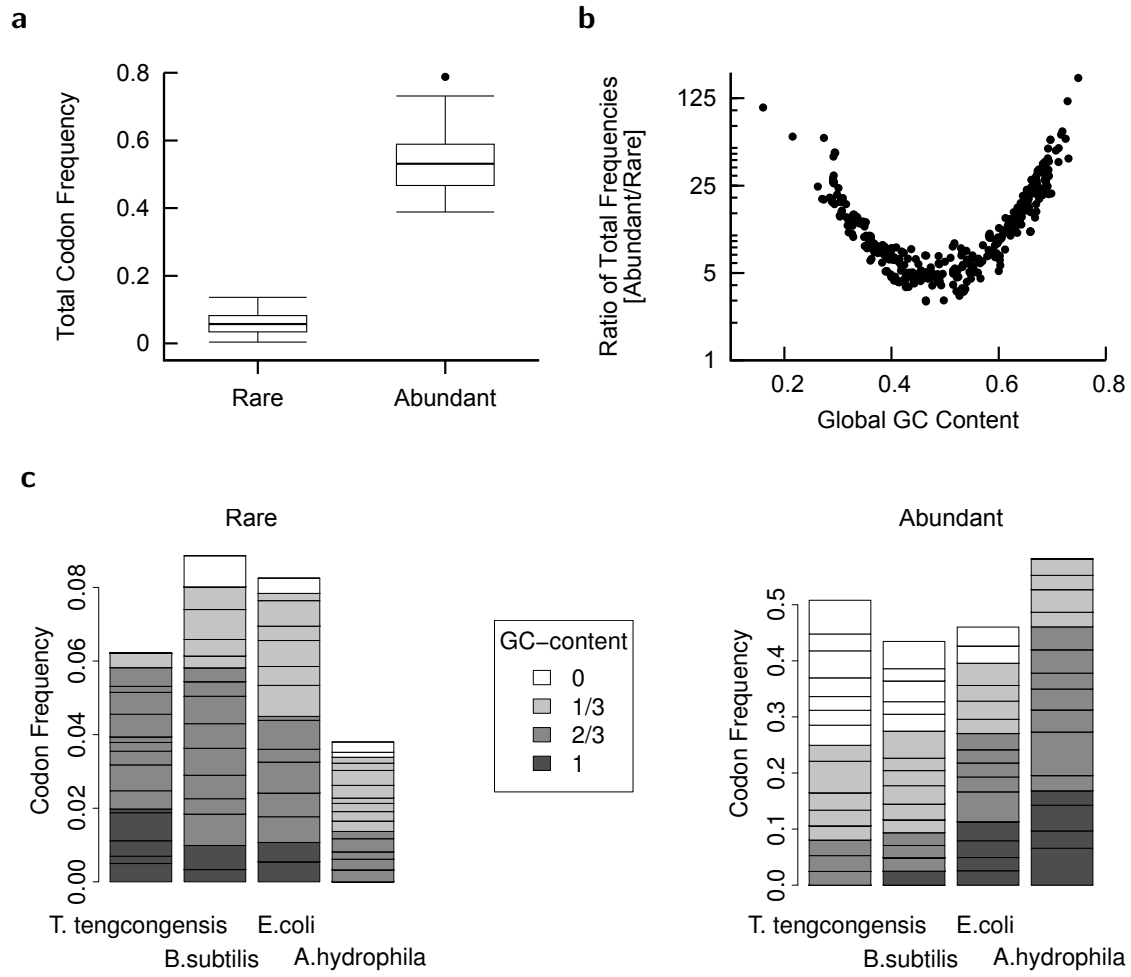


Figure 3.7.: Rare and abundant codons. (**a**) For each genome the 15 most rare and abundant codons were selected. We calculated the total frequency for both of these subsets, shown as a box plot. The two groups are clearly separated by almost an order of magnitude, the median total frequency of rare codons being ~ 0.06, that of abundant codons ~ 0.53. (**b**) As a measure for codon usage bias, we calculated the ratio of total frequencies of abundant over rare codons, shown as a function of global GC-content. The more extreme the GC-content, the stronger biased the codon usage, i.e. the most rare codons are found either in very GC- or AU-rich genomes. (**c**) The frequency distribution of rare codons and their GC-content, indicated by gray levels, is shown on the left for four different genomes. GC-content of the genomes increases from left to right. The higher the GC-content of the genome, the more rare GC-poor codons can be found and vice verse. On the right the frequency distribution and GC-content of abundant codons is shown. In contrast to rare codons, there are more GC-rich codons the higher the GC-content of the genome. Note that *E. coli* is rather balanced.

3.2.4. Properties of rare and abundant codons

The reduction of GC-content at the beginning of genes in *E. coli* might be a side effect of selection for rare and consequently slowly translated codons *per se* to prevent traffic jams of elongating ribosomes. To investigate this hypothesis, we defined subsets of codons based on how often they are used in a specific genome. We termed the 15 codons with lowest abundance rare, and those 15 with highest frequency abundant. With this definition, only about 5 per cent of all codons of an organism were on average rare codons, while about 50 per cent were abundant (fig. 3.7a). Interestingly, the GC-content strongly determines the overall codon distribution. In bacteria with a GC-content of 0.5, abundant codons were about 5-fold more frequent in the genome than rare codons, while bacteria with more extreme GC-content showed up to 100-fold difference in frequency (fig. 3.7b).

The GC-content of an organism is also reflected in the GC-content of rare and abundant codons in a particular genome (fig. 3.7c): In an AU-rich organism such as *T. tengcongensis*, rare codons are GC-rich, and abundant codons are AU-rich. Likewise, in GC-rich organisms, such as *A. hydrophila*, AU-rich codons are enriched among the rare codons, and abundant codons show high GC-content. In contrast, bacteria with intermediate GC-content show no particular selection for GC-rich or GC-poor codons. In these organisms rare codons are not biased towards a particular GC-content, and a selective pressure for the usage of rare codons *per se* is unlikely to strongly influence the local GC-content at the gene start.

3.2.5. Rare codons are selected to reduce GC-content in E. coli

Unusual codon usage in the first 10 codons is determined by a strong relative increase of rare codons in *E. coli* (fig. 3.8a), and a slight suppression of abundant codons. We next divided the sets of rare and abundant codons in two subsets: Those with G or C at the third position, and those with an A or U, termed GC3 and AU3 codons, respectively. This allowed us to discriminate between the two competing hypothesis for unusual codon usage: If there is a selective pressure to increase the frequency of rare codons downstream of the translation start in order to slow down early elongation ("ramp hypothesis"), we expect an increase of rare codons irrespective of their GC3-content. In contrast, if codons are chosen in order to disrupt the mRNA structure, we expect an asymmetry between GC3-rich and AU3-rich codons. Figure 3.8b unveils a clear asymmetry of AU3 and GC3 codons in the set of rare codons: Rare AU3 codons are enriched at the beginning of genes, while rare GC3 are not affected in their frequency. We also noted an asymmetry among the sets of abundant codons: Abundant GC3 codons are strongly depleted at the beginning of genes, while abundant AU3 codons are slightly enriched (fig. 3.8c). Depletion of abundant GC3 codons can only be partially explained by the choice of amino-acids (see null model). In summary, our analysis supports that rare codons are selected because of their GC-content and not because they are rare *per se*.

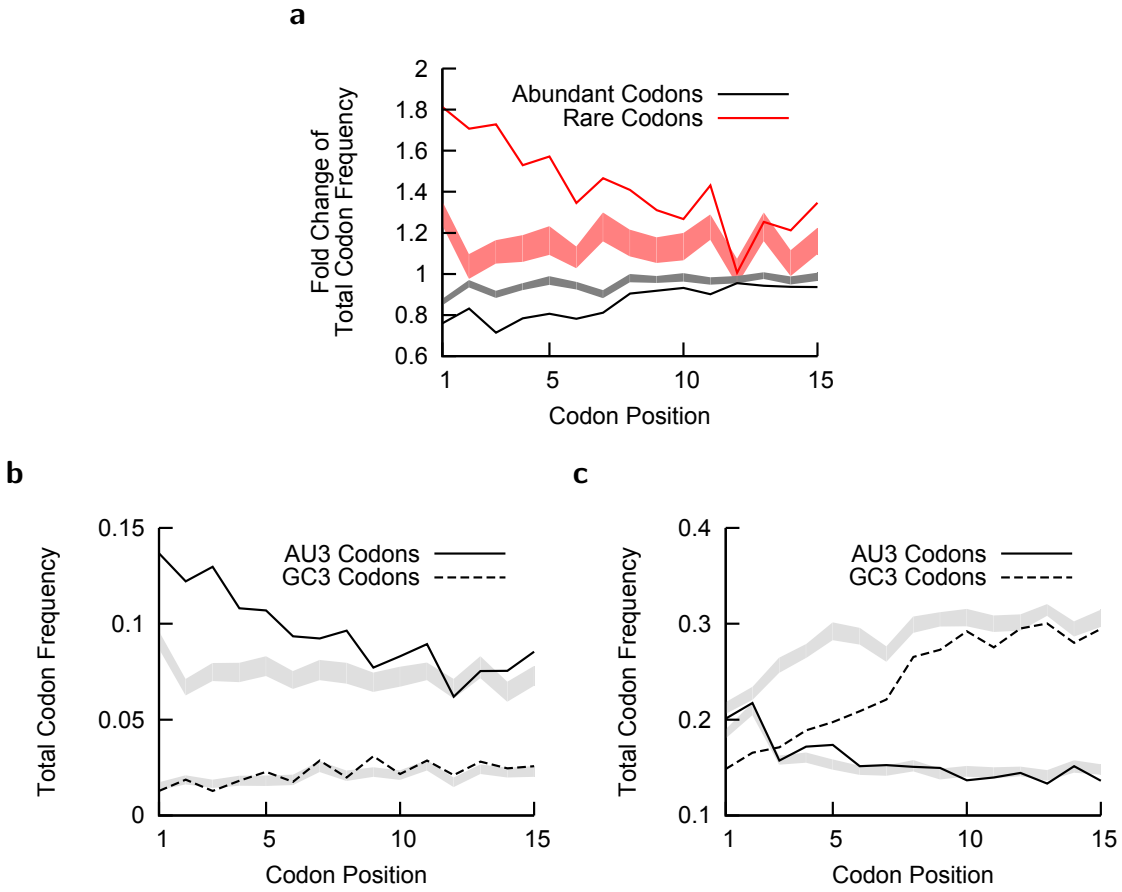


Figure 3.8.: Enrichment of extreme codons at beginning of genes in *E. coli*. (**a**) In *E. coli*, rare codons are enriched while abundant codons are depleted at the beginning of genes. Fold change was calculated with respect to the corresponding total codon frequency in the genome. The grayed areas show the average fold change ± standard deviation as obtained from the sequences with shuffled synonymous codons (SSC). (**b**) Rare codons are differentiated with respect to their third base, i.e. partitioned into AU3 and GC3 codons. Enrichment of rare codons can be attributed to an increased usage of AU3 codons, whereas GC3 codons are not enriched. Grayed areas show the average total frequency ± standard deviation estimated from the null model SSC. (**c**) The same plot as in (**b**) but for abundant codons. Frequency of GC3 codons is strongly reduced, leading to the observed pattern in panel (**a**) of decreased usage of frequent codons.

3.2.6. Wide-spread selection for reduced GC-content at gene start

If there is a generic pressure to reduce mRNA secondary structure which in turn determines a specific codon usage at the beginning of genes, the observed asymmetry in *E. coli* should be not an isolated exception but an example for a universal trend. We thus compared different genomes and asked whether selection of rare codons is compatible with the "ramp hypothesis" or with the "structure hypothesis". The "ramp hypothesis" predicts that rare codons are selected because they are translated with a lower efficiency, and accordingly

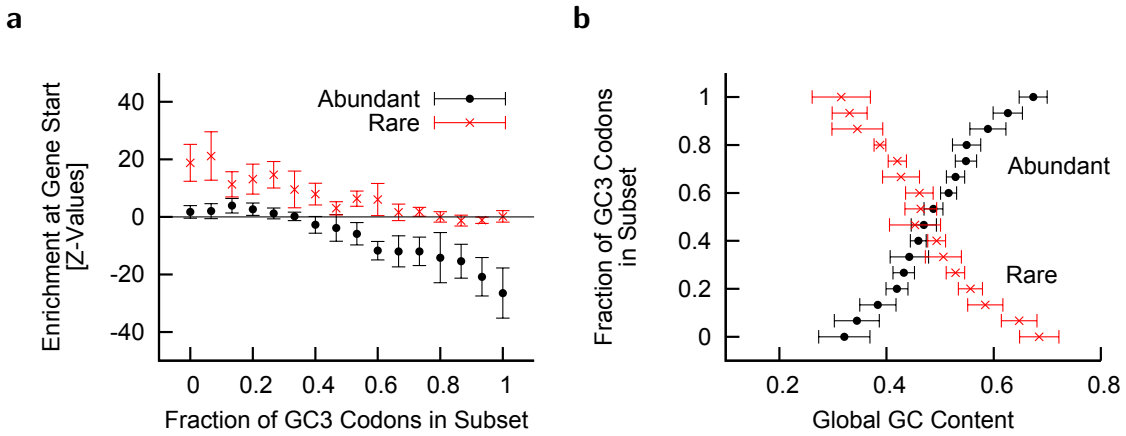


Figure 3.9.: Enrichment of extreme codons in bacterial genomes. Enrichment was assessed by calculating empirical Z values of fold change of total codon frequency at the beginning of genes for rare and abundant codons. To this end we used the null model of synonymous shuffled codons (SSC). (**a**) The fraction of GC3 codons in the subset of rare and abundant codons for each genome was calculated. The panel shows the mean enrichment $\pm$ standard deviation as a function of the fraction of GC3 codons. There is only an increase of rare codon and a reduction of abundant codon usage if it leads to a decrease of GC3-content. (**b**) The fraction of GC3 codons in the subset of rare and abundant codons is shown as a function of global GC-content. Averages $\pm$ standard deviations of global GC-content were calculated for subsets of rare and abundant codons having the same number of GC3 codons. The higher the global GC-content the more GC3 codons are found among abundant codons and correspondingly more AU3 codons are rare.

abundant codons are depleted at gene start. In contrast, the "structure hypothesis" predicts that rare codons would be enriched only if they are AU3-rich as these result in weaker structure, and likewise abundant only be depleted if that set is GC3-rich. When we grouped bacteria according to the GC3-content in their sets of rare or abundant codons, we observed that abundant codons are depleted only if the fraction of GC3 codons in that set is above 50%, and an enrichment of rare codons occurs, when rare codons have on average a AU3-content above 50% (fig. 3.9a). This further supports the "structure hypothesis". The fraction of GC3 codons among the rare and abundant codons is strongly determined by the GC-content of the genome. In GC-rich organisms, the set of rare and abundant codons is typically dominated by AU3 and GC3 codons, respectively. Correspondingly, the relation is reversed at the other end of the spectrum (fig. 3.9b). On ground of this observation the "structure hypothesis" thus predicts an enrichment of rare codons in a GC-dependent manner. Indeed, we find that rare codons are only enriched in organisms with a GC-content higher than ~ 0.45 (fig. 3.10a). This increased usage of rare codons is mirrored by a depletion of abundant codons.

Based on the relation between the global GC-content and the fraction of GC3 codons in the set of rare and abundant codons (fig. 3.9b), the two hypotheses make divergent predictions about the GC3-content at the beginning of genes. The "ramp hypothesis" predicts an

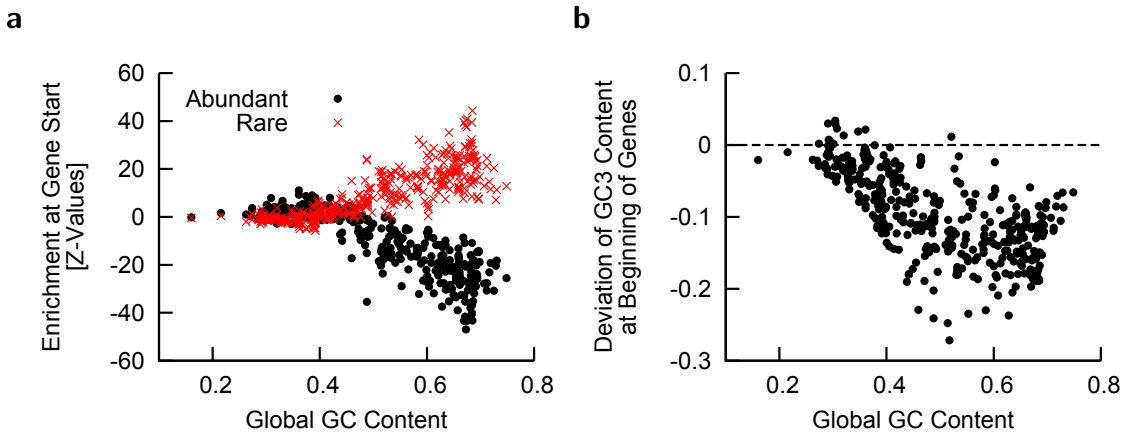


Figure 3.10.: Asymmetry of GC3-content at beginning of genes in different genomes. (**a**) Enrichment of extreme codons, quantified by Z values, is shown as a function of global GC-content. Rare codons are only enriched if the global GC-content is larger than about 0.45. Correspondingly, a reduction of the usage of abundant codons is observed. (**b**) The average deviation from the genomic GC3-content for codons 1 to 5 depends on the global GC-content. Virtually all genomes show a reduction in GC3-content at the gene start, and genomes with higher genomic GC-content typically show a stronger reduction (correlation coefficient $r = -0.66$).

enrichment of GC3 codons in AU-rich genomes at gene start, and a corresponding depletion of GC3 codons in GC-rich genomes. In contrast, an asymmetry is inferred from the "structure hypothesis": Only a reduction of the GC3-content at the beginning of genes is to be expected, with stronger depletion for genomes with higher GC-content. In the 414 analyzed genomes, the GC3-content behaves strongly asymmetric, as we observed a reduction in some genomes but hardly any increase in local GC3-content proximal to the start codon (fig. 3.10b). Importantly, the reduction in the local GC3-content is stronger for genomes with a higher GC-content suggesting that a reduction of the GC-content is one of the main driving forces behind the unusual codon bias just downstream of the start codon. It should be noted that all results are similar for archaea (fig. B.2 in appendix B.3) implying a universal rather than a domain-specific evolutionary trait.

3.2.7. Evolutionary simulations confirm that unusual codons are required to reduce secondary structure

We next sought to reproduce the unusual codon usage downstream of the gene start by evolutionary simulations, where genomes were optimized to have a reduced secondary structure at the gene start. We generated a randomized genome, where we shuffled synonymous codons, except for those coding for translation start and stop, within a gene. As a basis for the randomization, we used the genome of *E. coli*. We then evolved *in silico* this randomized genome towards less secondary structure around the start codon using an evolutionary op-

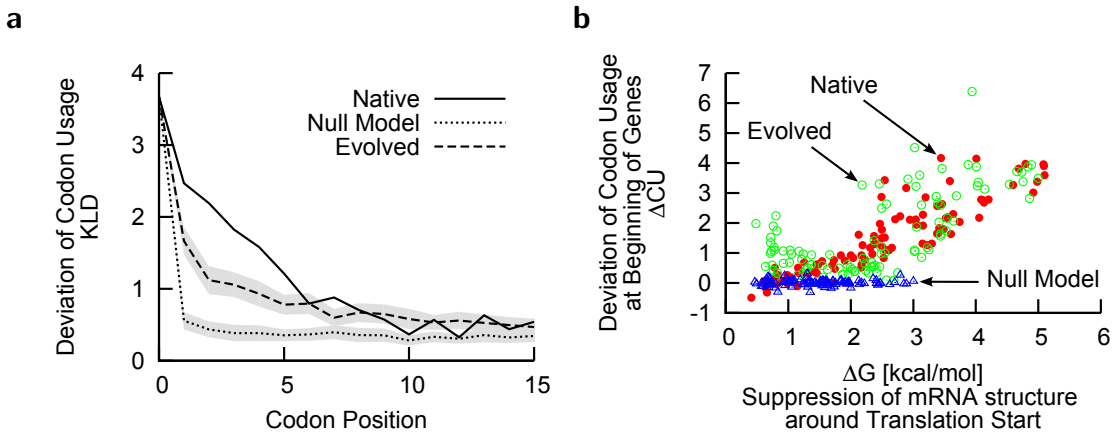


Figure 3.11.: Simulated evolution of codon usage near start codon. Synonymous codons of genes with non-overlapping sequences were shuffled, and subsequently optimized by an evolutionary algorithm to resemble native mean folding energy. (**a**) The KLD of initial, *in silico* evolved, and native sequences of the *E. coli* genome. The KLD of evolved sequences differs significantly from the null model (SSC) and shows similarity with the KLD calculated from *E. coli* sequences. (**b**) The correlation between deviation from codon usage, ΔCU and suppression of secondary mRNA structure, ΔG, was obtained by evolutionary simulations, initialized with SSC null model sequences, of 100 genomes. Note that ΔG was calculated slightly different here as for the determination of G_0 only sequences downstream of the start codon were taken into account.

timization algorithm (67). Within the algorithm, the mutator chooses synonymous codons with a probability distribution corresponding to the global codon usage. We ran 100 simulations with population sizes of 1000, and evolved the genome so that the energy profile of each gene around the start codon is similar to the average energy profile in the *E. coli* genome. Similarly to the *E. coli* genome, the local codon usage around the start codon deviates from global codon usage (fig. 3.11a). Quantitatively we even observed a slightly stronger deviation from the global codon usage in the *E. coli* genome than in our simulation which might be due to additional evolutionary pressures that are not present in our simulations, including the avoidance of secondary structures on a range longer than the 39 nucleotides used in the simulation. Furthermore, we observed that the secondary structure reduction correlated with a decrease in GC3-content. However, the deviation was smaller when compared to native *E. coli* sequences (fig. B.3 in appendix B.3).

But can evolutionary simulations recover the strong correlation between unusual codon usage and suppression of secondary structure? In an attempt to simulate organisms with different GC-content and different evolutionary pressure to reduce the secondary structure, we selected a representative sample of 100 bacteria. Synonymous codons were shuffled within each gene of these 100 genomes. The shuffling of codons did not introduce any significant deviation of codon usage at gene start (i.e. $\Delta\text{CU} \approx 0$, see fig. 3.11b). However, the secondary structure downstream of the gene start is still repressed to some extent as also

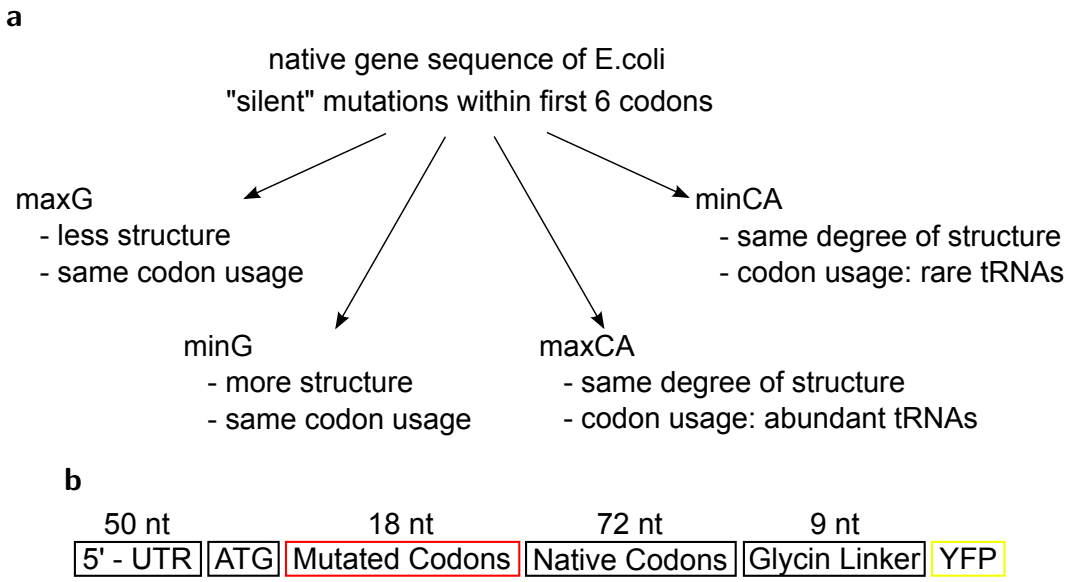


Figure 3.12.: Experimental strategy for disentangling codon usage and folding energy. (**a**) We selected genes encoding for cytoplasmic proteins for which we could vary codon usage and folding energy most efficiently. By introducing silent mutations in the first six codons after the initiation codons we generated all synonymous sequences *in silico*. These synonymous sequences were computationally evaluated with respect to local codon usage and folding energies. From all these synonymous sequences the extreme cases were taken for further investigation. (**b**) We cloned the selected sequences into a plasmid carrying the yellow fluorescent protein gene.

non-synonymous codon usage seems to be biased towards less structured RNA. After subsequent evolutionary simulations the suppression of secondary structure is similar to that of native sequences of the corresponding bacterium. The evolved sequences exhibit similar deviation in the codon usage with a correlation between ΔCU and ΔG similar to the existing bacterial genomes (fig. 3.11b). Taken together, these evolutionary simulations show that an evolutionary pressure to reduce secondary structure around the start codon leads to an unusual codon usage and a reduction of the GC3-content at the beginning of coding regions.

3.2.8. Experiments confirm strong effect of folding on translation efficiency

We interpret suppression of mRNA folding around translation start as a necessity to keep the ribosome binding site free of structure allowing efficient binding of ribosomes, thus being an important determinant of translation efficiency. To address this hypothesis, we dissected experimentally the role of the codon usage and mRNA secondary structure for translation efficiency. We selected computationally genes from *E. coli* for which the folding energy and codon usage can be varied independently (fig. 3.12). We found that for the genes *pykA* and *ypdE* these criteria were met best and therefore chose them for further investigation. For *pykA* and *ypdE* genes, we selected sequences which differ strongly in folding energy (*minG*

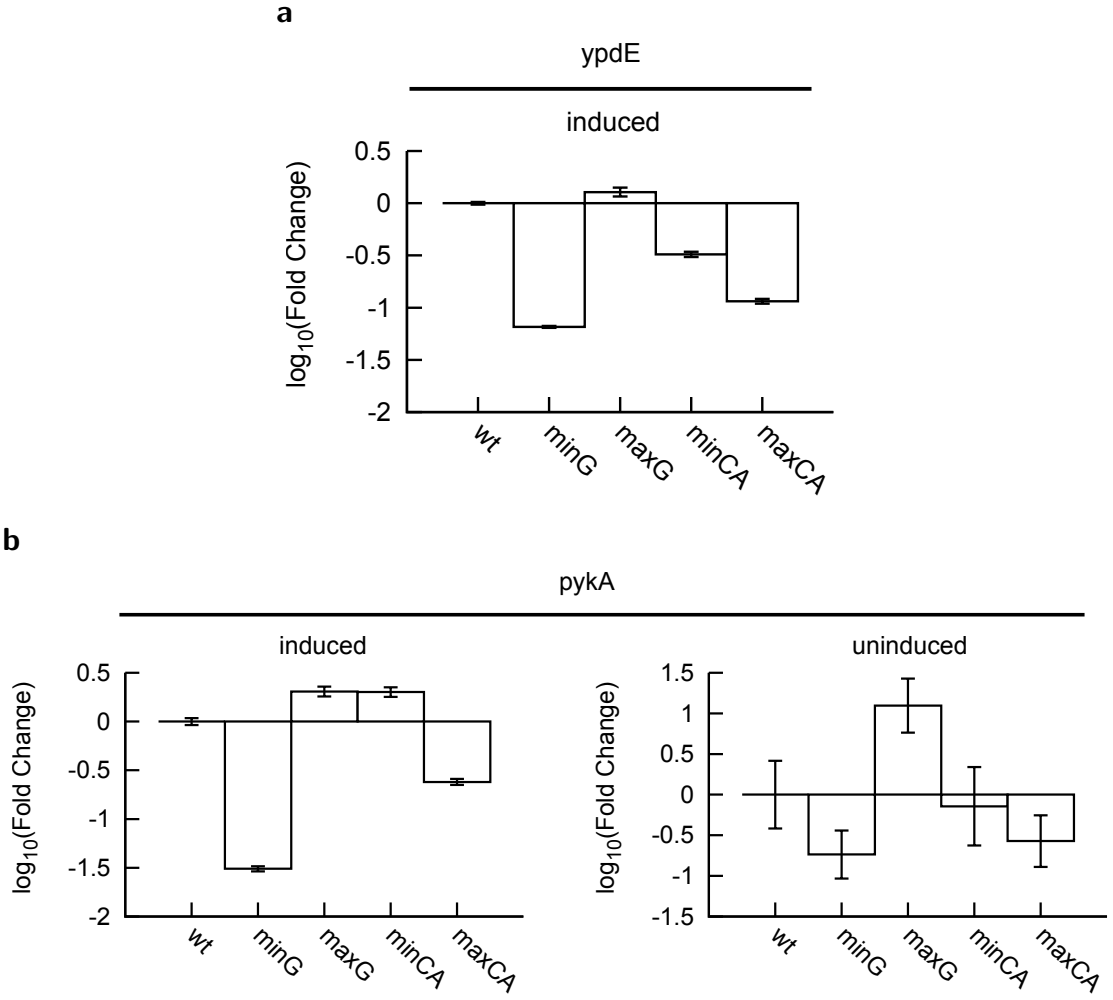


Figure 3.13.: Influence of synonymous mutations on translation efficiency. Constructs with various codon usage and folding energy were derived from two different *E. coli* genes (*ypdE* in (**a**) and *pykA* in (**b**)). They were expressed in *E. coli* in triplicates and the fluorescence was measured after and before induction with 45 μM IPTG by flow cytometry. The median of fluorescence distributions was averaged and normalized to wild-type expression. Errors based on the standard deviation of the median were calculated by propagating uncertainties. Changing mRNA folding energy and leaving codon usage unaltered had a pronounced and reproducible effect on translation efficiency. Sequences with strong secondary structure (*minG*) were much weaker expressed than wild-type (*wt*), whereas constructs with loose structure (*maxG*) showed increased expression. Effects on gene expression by modifying codon usage, i.e. maximal and minimal adapted, but not changing folding energy, were present but less pronounced and inconsistent (compare *minCA* and *maxCA* for usage of codons corresponding to rare and abundant tRNAs, respectively). Moreover effects of altering codon usage were less predominant for *pykA* before induction.

and *maxG* for minimal and maximal folding energy, respectively), but have similar codon usage as the native sequence (*wt*) to investigate the effect of mRNA secondary structure on translation (figs. 3.12a and B.4 in appendix B.3). Likewise, to study the effect of codon

usage we chose sequences that differ in the codon adaptation, but have similar estimated folding energy (*minCA* and *maxCA*, for minimal and maximal codon adaptation, respectively). These sequences were fused upstream of the coding region of a yellow fluorescent protein gene (fig. 3.12b), expressed in *E. coli*, and the translation efficiency was assessed by quantitative measurements of mRNA and fluorescence levels. Alterations in codon usage or secondary structures had no influence on mRNA abundance as confirmed by qRT-PCR (fig. B.5 in appendix B.3). In contrast, the protein abundance differed largely (figs. 3.13 and B.6 in appendix B.3); the observed differences in protein expression between the constructs were 20- and 60-fold for *ypdE* and *pykA*, respectively. For both genes, we found that the constructs with minimal folding energy (i.e. strongest secondary structure) showed by far the lowest expression, whereas constructs with maximal folding energy yielded the highest protein expression (fig. 3.13), as predicted by our theory. Codon usage also influenced the protein expression, but the effects were much weaker and rather inconsistent. For *ypdE*, the constructs with minimal and maximal codon adaptation had a reduced protein expression (fig. 3.13a). In contrast, for *pykA* we observed an enhanced expression for minimal, and a decreased expression for maximal codon adaptation (fig. 3.13b). Taken together these results clearly suggest that mRNA structure determines the translation yield whereas the effects of codon usage are less obvious. The induction of *pykA* resulted in very high protein levels, which may impact on cell physiology in general (fig. B.6b). We thus measured the protein levels of uninduced constructs. While the strong influence of folding energy was preserved, the construct with minimal codon adaptation showed no difference to the wildtype construct (fig. 3.13b).

3.2.9. The impact of slow codons at beginning of ORFs

Since rare codons in general correspond to low abundant tRNAs, their presence might have an influence on the early elongation and translation initiation rate. Slowly elongating ribosomes could prevent free ribosomes from initiating, since they occupy part of the ribosome binding site. Similar to Bulmer (15), we developed a simple model taking this effect into account by differentiating between early and late elongation, each taking place with a rate v_i^e and v_i^l, respectively. The reaction scheme thus reads

$$\mathrm{m_i^f} + \mathrm{R} \xrightarrow{k_i} [\mathrm{R_{in}}|\mathrm{m_i}] \xrightarrow{v_i^e} [\mathrm{R_{el}}|\mathrm{m_i}] + \mathrm{m_i} \xrightarrow{v_i^l} \mathrm{p_i} + \mathrm{R}. \tag{3.4}$$

Ribosomes R bind to the free ribosome binding sites forming an initiation complex $[\mathrm{R_{in}}|\mathrm{m_i}]$. During early elongation the ribosome leaves the initial state with a rate $v_i^e[R_{\mathrm{in}}|m_i]$, upon which the ribosome elongates further with a rate constant v_i^l. The corresponding set of

equations is given by

$$\frac{d}{dt}[R_{\text{in}}|m_i] = k_i m_i^f R - v_i^e[R_{\text{in}}|m_i] \tag{3.5}$$

$$\frac{d}{dt}[R_{\text{el}}|m_i] = v_i^e[R_{\text{in}}|m_i] - v_i^l[R_{\text{el}}|m_i] \tag{3.6}$$

$$m_i = m_i^f + [R_{\text{in}}|m_i] \tag{3.7}$$

$$R^T = R + \sum_k \left([R_{\text{in}}|m_k] + [R_{\text{el}}|m_k]\right), \tag{3.8}$$

where

$$v_i^e = \frac{1}{\sum_{j=1}^{M_i} v_{ij}^{-1}} \tag{3.9}$$

$$v_i^l = \frac{1}{\sum_{j=M_i+1}^{L_i} v_{ij}^{-1}}. \tag{3.10}$$

We apply again a quasi-steady state approximation

$$k_i(m_i - [R_{\text{in}}|m_i])R - v_i^e[R_{\text{in}}|m_i] \approx 0 \tag{3.11}$$

$$v_i^e[R_{\text{in}}|m_i] - v_i^l[R_{\text{el}}|m_i] \approx 0, \tag{3.12}$$

where we used the conservation relation (3.7) for mRNA but neglected that for ribosomes. The solution yields

$$[R_{\text{in}}|m_i] = m_i \frac{k_i R}{k_i R + v_i^e} \tag{3.13}$$

$$[R_{\text{el}}|m_i] = m_i \frac{v_i^e}{v_i^l} \frac{k_i R}{k_i R + v_i^e}, \tag{3.14}$$

and the equation governing protein synthesis therefore now reads

$$\frac{d}{dt}p_i = m_i v_i^e \frac{k_i R/v_i^e}{k_i R/v_i^e + 1} - \gamma_p p_i. \tag{3.15}$$

In the limit $k_i R/v_i^e \ll 1$, i.e. initiation is still much slower than early elongation, we obtain the old result eq. (2.19). However, for the other extreme $k_i R/v_i^e \gg 1$, the synthesis rate saturates and depends on the early elongation rate and mRNA concentration only

$$\frac{d}{dt}p_i = m_i v_i^e - \gamma_p p_i \qquad k_i R/v_i^e \gg 1. \tag{3.16}$$

This occurs when the ribosomes are blocking subsequent initiation, thus reducing the overall translation rate.

3.3. Discussion

3.3.1. Codon usage at beginning of genes is shaped by suppression of mRNA structure

Synonymous codons are not chosen randomly. The frequency of synonymous codons within the first codons of coding sequences deviates from the overall codon usage in a genome. Here we present a new aspect that drives the selection of unusual codons: Codons at the beginning of genes are selected to suppress mRNA secondary structure around the ribosome binding site. As reported in an earlier study (50), the pressure to disrupt such structures is stronger if mRNA tends to fold more stably, which is predominantly determined by the genomic GC-content. We confirmed these findings and linked them to the observed unusual codon usage at the gene start across many genomes. Both, suppression of mRNA folding and unusual codon usage, were found to be strongly correlated. More specifically, in genomes with a high genomic GC-content AU3 codons are found more frequently at the beginning of genes than elsewhere in the genome. This is consistent with the suppression of mRNA folding which is achieved by decreasing the local GC-content and also explains why we observe and enrichment of rare codons at the gene start: In a genome with high GC-content AU3 codons are on average less frequent, thus the subsets of rare and AU3 codons do overlap to a large extent. It becomes therefore evident that the feature a codon is selected for at the gene start is its GC-content and most likely not its genomic frequency and consequently its translation efficiency. Hence it is not surprising that there is no correlation between RNA structure and the usage of rare codons on the single genome level (23).

Previously rare codons at the gene start were proposed to slow down early elongation preventing ribosome cluttering later in the translation when the elongation speed increases (154). This theory would predict that the unusual codon usage is universal among bacteria and would not depend on specific genomic features of the organism like GC-content. In many organisms elongation speed might be slower at the beginning of genes, like in yeast (62, 155), but our analysis suggests that this is most likely a consequence of the need to suppress mRNA structure and the resulting use of rare codons. Moreover, our analysis of a simple mathematical model showed that slowing down early elongation might actually be harmful, since it could counteract an efficient initiation thereby reducing the overall translation rate. If necessary at all, jamming could be alternatively prevented by modulating the initiation rate. This would also prevent the sequestering of ribosomes at the beginning of genes, which might be beneficial due to a more economic employment of an important cellular resource. Consequently, it is not to be expected that there is a selective pressure for rare codons *per se*, but an enrichment of those which lead to a weaker folding of mRNA around the ribosome binding site.

3.3.2. Reduced mRNA folding is important for efficient translation initiation

Several studies show that strong secondary structures at the ribosome binding region reduce or even terminate translation initiation (30, 76). Here, we experimentally show that changing the folding energy while keeping the same codon usage at the beginning of native *E. coli* genes markedly affects translation efficiency. In contrast, alterations of the codon usage only while maintaining the same folding energy led to less conclusive results. Our approach to investigate the effects of codon usage and folding at the gene start differs markedly from an earlier study (76). Kudla et al. constructed a library of green fluorescent proteins with synonymous substitutions of codons along the whole transcript. They found that the degree of secondary structure around the translation start explained expression levels best. In addition they made use of a 28-codon tag fused to the 5′ terminus of 72 GFP constructs. This tag featured weak mRNA secondary structure and low codon adaptation and produced consistently high expression. In that study effects of codon usage at the beginning of the genes might be occluded by differentially substituted codons further downstream. In contrast, we only changed the codons at the very beginning of genes. Hence we clearly can attribute differences in the translation efficiency to these few codons. Furthermore, we used native *E. coli* sequences fused to a reporter gene instead of *gfp* alone, which might be not well adapted to the endogenous translational apparatus.

Both, ours and the previous study (76), show that translation efficiency is strongly modulated by the folding energy in the initiation region. In addition our results suggest that native gene sequences might be already rather optimized in terms of mRNA structure. Indeed, it was much easier to shut down translation by stabilizing mRNA folding than to increase translational efficiency, which is in line with earlier findings (30). This holds for both investigated genes, although their absolute expression levels differed significantly. Moreover, the increase and smaller spread of folding energies around the translation start in *E. coli* suggest that suppression of mRNA structure is a necessary condition for efficient translation of a gene. It might thus not be surprising that no direct correlation between gene expression levels and the folding was found (155), since there are additional factors, like strength and position of the Shine Dalgarno sequence (124), contributing to the overall translational efficiency of a gene.

3.3.3. Conclusion

Our analysis shows that the evolutionary pressure to keep the ribosome binding regions free of structure is very strong. Surprisingly, in *E. coli* proteins show even an enrichment of amino acids at the N-terminus that are encoded by codons with A or U at the first or second nucleotide position. The role of the codon usage in defining mRNA structure might be not only restricted to the translation start. Codon choice along the coding sequences may also be shaped by various requirements towards specific secondary structures. For

example, mRNA stability, micro-RNA binding or RNA-binding proteins may require certain structures, which would lead to the choice of specific codons. Thus, evolution has to solve a multi-dimensional problem: While efficient elongation and error-reduction favor the usage of frequent codons, certain structural requirements for the mRNA may need infrequent codons. Furthermore, rare codons coordinate processes downstream of translation, including co-translational folding (169, 74). Thus, codon usage is shaped by many possibly opposing constraints which we are just beginning to understand.

4. Translational coupling and chemotaxis efficiency

This chapter is an extended version of "Role of translational coupling in robustness of bacterial chemotaxis pathway" (88). Contributing authors: Linda Løvdok (Experiment), Kajetan Bentele (Theory and Simulation), Nikita Vladimirov (Bioinformatics Analysis), Anette Müller (Experiment), Ferencz S. Pop (Experiment), Dirk Lebiedz, Markus Kollmann, and Victor Sourjik.

4.1. Introduction

Many genes in *E. coli* and other bacteria are organized in operons (120, 71, 66), implying the co-transcription of several genes as a single messenger RNA (mRNA). In such polycistronic mRNAs, the start codon of an open reading frame (ORF) is often located near the stop codon of an upstream ORF. Consequently, once ribosomes have translated a gene they may re-initiate translation of the downstream gene. However, it is also conceivable that ribosomes terminating at the upstream gene influence the structure of the mRNA around the ribosome binding site (RBS) of the downstream ORF. As we have seen in the preceding chapter, there is an evolutionary pressure to keep the RBS of a gene free of secondary structure likely to allow efficient translation initiation. In addition, terminating ribosomes could facilitate the unfolding of the mRNA structure around the translation start site of the downstream gene thus enhancing de-novo initiation. These mechanisms may thus couple translation of adjacent ORFs in an operon. Translational coupling is defined as the interdependence of translation efficiency of neighboring genes encoded by the same polycistronic mRNA. Such coupling helps to maintain a constant ratio of proteins levels and might thus be beneficial in ensuring proper stoichiometry of subunits of protein complexes or enzymes of a metabolic pathway and has been previously described in *E. coli* (8, 85, 109, 126).

To investigate translational coupling and its possible beneficial effects, we need to employ a system with a well understood mapping between genotype and phenotype. In addition, we have to be able to assess the fitness of a given genotype. The chemotaxis pathway of *E. coli* represents such a model system as it is one of the best studied prokaryotic systems for signaling and robustness and meets all these criteria (159, 138). Here we present experimental evidence that translational coupling indeed exists between pairs of chemotaxis genes. Such coupling provides a mechanism to reduce the negative impact of noise on cellular functions as

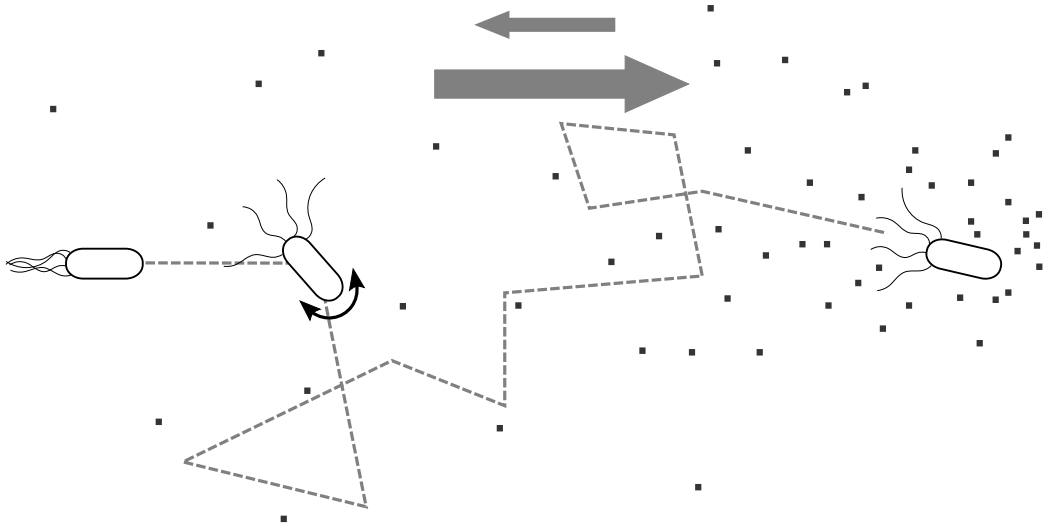


Figure 4.1.: Chemotaxis strategy of *E. coli*. Cells perform a random walk in the absence of chemical gradients: Swimming is interrupted by tumbling which causes a random reorientation. However, tumbling is suppressed, i.e. the frequency of tumble evens is decreased, if cells swim into favorable directions. This behavior results in a biased random walk towards higher attractant concentrations.

shown experimentally and corroborated further by our theoretical analysis. More specifically, our analysis suggests that operon organization is driven by selective coupling of specific chemotaxis genes. In the following we briefly describe the search strategy of *E. coli* and how it is implemented at the molecular level. Finally we discuss important features of the chemotaxis pathway related to its molecular organization.

Bacterial chemotaxis is the directed movement along chemical gradients. Many bacteria have evolved a sensory system and a motile apparatus allowing them to move towards higher concentration of attractant chemicals and to avoid toxins (repellents). Chemotaxis is achieved by *E. coli*, the best studied model system for chemotaxis of bacteria, by a biased random walk (fig. 4.1). Smooth runs, corresponding to a counterclockwise (CCW) rotation of flagellar motors are interrupted by a clockwise (CW) motor rotation causing the cell to stop and tumble thereby randomizing the direction of the next run. Cells swimming in favorable directions suppress tumbling, i.e. they reduce their tumbling frequency, and correspondingly prolong running. This results in a net drift towards regions of higher attractant concentration. Conversely, cells heading towards higher repellent concentration tumble more frequently, and thus tend to change their direction thereby increasing their chance to escape harmful chemicals (12, 11, 1). Hence the cells react to temporal changes of ligand concentrations along their swimming path, as they are not able to sense directly the spatial gradient due to their small size.

Diverse transmembrane receptors, called methyl-accepting chemotaxis proteins or MCPs, enable *E. coli* to detect a variety of amino acids, sugar and dipeptides, in addition to pH,

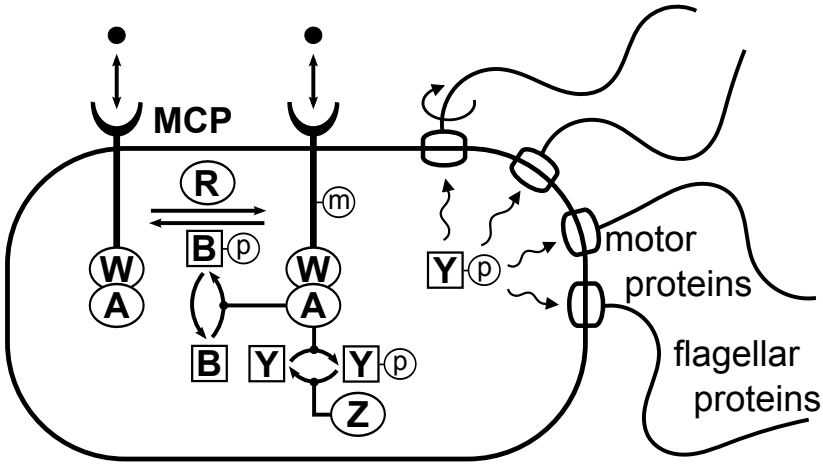


Figure 4.2.: Chemotaxis pathway in *E. coli*. Transmembrane receptors (MCP) sense the chemical environment and transmit signals by conformational changes into the cell. The cytoplasmic part of the receptor forms a complex with the adaptor protein CheW (W) and the protein kinase CheA (A). CheA phosphorylates itself with a rate dependent on the receptors' activity state. The phosphoryl groups (p) are transferred to CheY (Y), a small diffusable messenger protein. The phosphorylated CheY in turn interacts with the flagellar motors and induces tumbles. The phosphoryl groups are constantly removed by the phosphatase CheZ (Z).
Attractant binding inactivates the receptors. In turn autophosphorylation of CheA stalls causing a drop of phosphorylated CheY. Consequently, tumbling is suppressed thereby prolonging swimming into the favorable direction of increasing attractant concentrations. Adaptation to the increased ligand concentration is provided by methylation of receptors which raises activity levels of receptors and allows cells to detect further concentration changes. A pair of counteracting enzymes, CheR (R) and CheB (B), add and remove methyl groups (m). The demethylating activity of CheB is greatly enhanced upon phosphorylation through CheA. Figure adapted from (3).

temperature and redox state (138). Strongly expressed receptors as those for serine (Tsr) and aspartate (Tar) are present in high copy numbers with 24×10^3 to 37×10^3 receptors per cell in *E. coli* wild type cells grown in minimal medium. In contrast, low abundant receptors, as those for sensing ribose and galactose (Trg), only number several hundred (83). Signaling in the chemotaxis pathway relies on the phosphorylation of diverse proteins (fig. 4.2), with the histidine kinase CheA being the key enzyme (57, 159, 6). Dimers of receptors form on their cytosolic portion together with CheA and the adaptor protein CheW large clusters at the cell poles (138). Activity of these receptors is modulated by extra-cellular ligand concentrations which in turn regulates autophosphorylation activity of CheA (141, 151). Signaling from the receptors to the motors is mediated by a diffusable response regulator, CheY, which is rapidly phosphorylated by CheA (140). The levels of phosphorylated CheY control the direction of the flagellar motor proteins and thus the swimming behavior of the bacterium (fig. 4.3) (2, 24). Levels of phosphorylated CheY are additionally controlled by a phosphatase CheZ (14, 171). The adaptation to external stimuli takes place on a slower time scale and

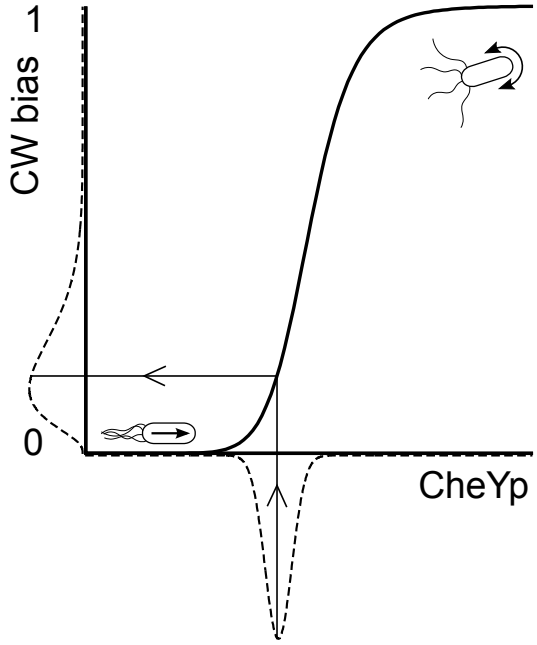


Figure 4.3.: Phosphorylated CheY controls the CW bias of the flagellar motors. An increase of attractant concentrations causes a drop of CheYp and thus a decreased CW bias. As a consequence cells continue to swim into the chosen direction. On the contrary, falling attractant concentration lead to an enhanced phosphorylation of CheY in turn increasing the CW bias. This way cells pick randomly a new direction increasing their chance to swim in a more favorable direction afterward. In the adapted state cells exhibit a CW bias around 0.2. The motor-response curve is extremely steep with a measured Hill coefficient of 10 (24). Thus already small variations of the CheYp level cause fluctuation of the CW bias over a wide range. Thus a strong evolutionary pressure to suppress fluctuations of the adapted pathway output is to be expected.

is controlled by two counteracting receptor modification enzymes, CheR and CheB, which add and remove methyl groups. They both adjust the activity of the chemotaxis receptors by changing the receptors' methylation state, thereby restoring the pathway's output, the concentration of phosphorylated CheY, to its pre-stimulus value (141, 80, 159). As this process is slower than the initial response, the methylation state of the receptor provides a memory about past conditions experienced by the cell (138).

The performance of the signaling system is paramount: *E. coli* can detect concentration changes which amount to less than 10 molecules per cell volume and exhibits a signal gain by a factor of ~ 100 (127, 92, 138). Thanks to the methylation system cells retain, at least for some attractants, sensitivity in ambient concentrations spanning five orders of magnitude (138). The astonishing signal amplification is most probably achieved in two steps: First, receptors forming large clusters and receptor-kinase complexes were shown to account for 35-fold amplification of chemotactic signals (141). Second, the motor response curve (fig. 4.3), translating the level of phosphorylated CheY into the CW bias, is extremely steep

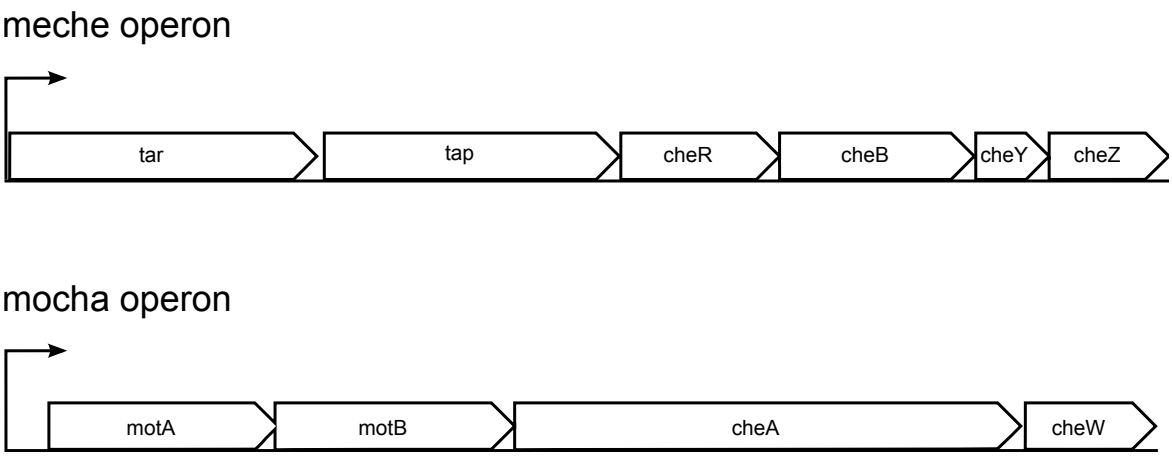


Figure 4.4.: The *meche* and *mocha* operons. The receptor genes *tar* and *tap* are transcriptionally coexpressed along with genes important for adaptation to constant stimuli, *cheR* and *cheB*, and the genes coding for the response regulator CheY and its phosphatase CheZ. The *mocha* operon harbors genes coding for the flagellar motor proteins, MotA and MotB, the adaptor protein CheW and the kinase CheA. Scheme and annotation adapted from the EcoCyc database (71).

with a Hill coefficient of 10 (24). When these two amplification steps are combined, they are sufficient to explain the observed gain (141, 138).

The pathway does not only perform well in terms of signal amplification but has also become a paradigmatic model for robustness of signaling networks. It was shown that robust adaptation independent of parameters is a key feature of this signal transduction pathway which could be explained by the assumption that the methylesterase CheB preferentially demethylates active receptors (7, 3, 99, 166). However, bacterial cells are not only exposed to a fluctuating environment. As we have seen in chapter 2, there are intercellular variations of protein levels due to noisy gene expression which might interfere with the proper functioning of the pathway. As noted earlier, cells exhibiting a clockwise motion of their flagellar proteins tumble, those with counterclockwise are swimming into one direction. Thus the adapted level of CheYp has to be tightly controlled despite fluctuating protein levels, otherwise cells would swim straight into one direction or tumble all the time and therefore would be unable to perform chemotaxis (fig. 4.3). The topology of the chemotaxis pathway allows for the compensation of correlated gene expression noise, relying on the opposing enzymatic activities of CheA/CheZ and CheR/CheB (73). Such correlated fluctuations can be achieved by coupling of gene expression. Indeed, chemotaxis proteins are organized in two operons, the *meche* operon which encodes two receptor proteins, Tar and Tap, along with proteins constituting the signaling network, CheR, CheB, CheY and CheZ and the *mocha* operon encoding CheA, CheW and the flagellar motor proteins (fig. 4.4) (103). This organization of the genes implies coupling on the transcriptional level and it has become apparent that such clustering is the strategy selected by evolution to assure the correlation of protein levels.

However, the reason for the order of genes within the operons remained unresolved. Translational coupling correlates expression of adjacent genes and is therefore a promising candidate to account for selection of a specific gene order. The chemotaxis pathway is well suited to investigate this mechanism: First, the constituents of the pathway as well as their interactions have been thoroughly investigated and are known in great detail. Second, chemotaxis is under strong selection as it enables bacteria to search for optimal growth conditions thereby conferring a competitive advantage. Hence it is plausible that evolution also took advantage of translational coupling in order to further optimize the functioning of the pathway. Third, we can asses the fitness of a population by quantifying the variations of the adapted tumbling frequency. This is an appropriate measure for fitness, as these fluctuations have to be kept small for efficient chemotaxis.

Next we present experimental evidence that translational coupling between adjacent genes of the *meche* and *mocha* operon indeed exists. Thereafter we show that pairwise coexpression of genes improves chemotactic performance indicating the possibly beneficial impact of translational coupling on the cellular fitness. A model of the chemotaxis pathway and translational coupling of gene expression is introduced. Both these models are used to investigate *in silico* the effect of gene ordering within the *meche* operon on the robustness of the pathway output under noisy gene expression.

4.2. Results

4.2.1. Translational coupling between chemotaxis genes

We first tested whether neighboring chemotaxis genes are coupled on the translational level by analyzing four pairs of genes: *cheR_cheB*, *cheB_cheY*, and *cheY_cheZ* from the *meche* operon and *cheA_cheW* from the *mocha* operon. The pairs were cloned as found in the genome with the second gene fused to a enhanced yellow fluorescent reporter gene (*eyfp*) as shown in fig. 4.5a. The translational level of the first gene was selectively varied by placing ribosome binding sites of different strength upstream of the translation start codon. To test the efficiency of these RBS, *eyfp* was fused to the first gene of the pairs and expressed from a monocistronic control construct (fig. 4.5a). Constructs with a stronger RBS showed a translational enhancement which varied from five to nine (fig. 4.5b). The values of upregulation at varying (0 to 50 μM) levels of IPTG induction did not differ significantly and were thus averaged. For CheA this analysis was complicated by the fact that there exists two alternative forms. The short form $\mathrm{CheA_S}$ is expressed from an open reading frame which starts 291 bases downstream of the translation start site of the long form, $\mathrm{CheA_L}$ (136). Hence, changing the first RBS had only a moderate effect on the overall expression level of CheA. Consequently, we tested the translational enhancement using another strategy by comparing constructs expressing $\mathrm{CheA_L}$ under the altered RBS and $\mathrm{CheA_S}$ under the native RBS with those expressing only $\mathrm{CheA_S}$ under the upregulated RBS. The net level

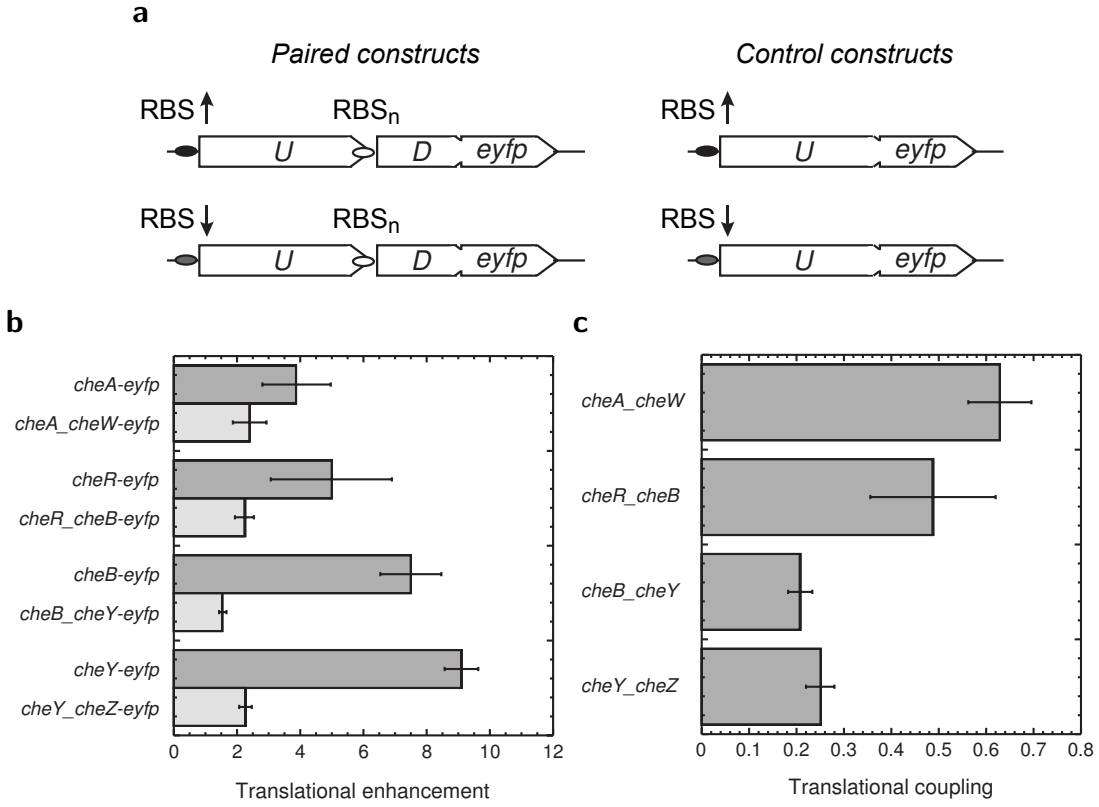


Figure 4.5.: Translational coupling between neighboring genes. (**a**) Experimental strategy. Bicistronic constructs that contained pairs of neighboring chemotaxis genes in their chromosomal arrangement (U, upstream gene; D, downstream gene) were cloned under RBS of different strength to create a C-terminal YFP fusion to a downstream gene. Strong RBS is indicated by a black oval and an up arrow, weak RBS by a gray oval and a down arrow. As a control of the RBS strength, the same sequence was placed in front of the monocistronic YFP fusion to the upstream gene. The downstream gene is under control of its native RBS (RBS_n, open oval). Expression was analyzed using FACS as described in the appendix C.1. (**b**) Direct (dark-gray) and indirect (light-gray) up-regulation of expression level of the fusion reporter by the stronger RBS, defined as the ratio of expression of constructs with the strong RBS to expression of corresponding constructs with the weak RBS. (**c**) Translational coupling, defined as the ratio of indirect to direct up-regulation of expression levels by the stronger RBS. Error bars in (**b** and **c**) indicate standard deviations. Figures with adapted captions are taken from (88).

of translation of $CheA_L$-YFP and $CheA_S$-YFP in the first construct was about four times higher than of $CheA_S$-YFP in the second construct. For the *cheA_cheW* pair, translation was regulated by using constructs that express either only short version of CheA or both long and short versions.

An increased translational efficiency of the first gene in each pair led to an elevated expression of the downstream gene in all cases indicating the existence of translational coupling. This indirect upregulation increased expression by about a factor of two for all constructs

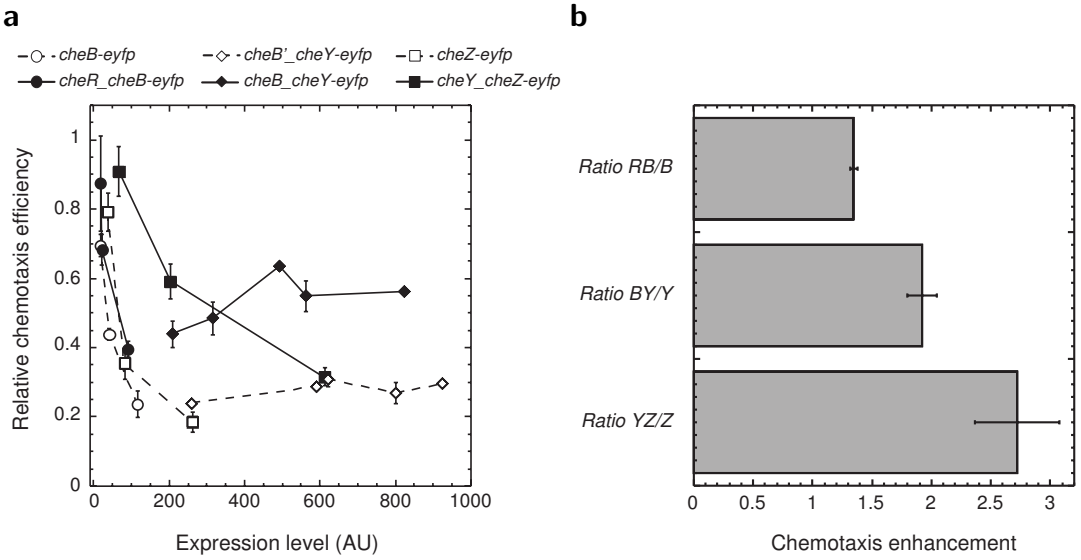


Figure 4.6.: Improvement of chemotaxis by coexpression of signaling proteins. (**a**) Dependence of the chemotaxis-driven spreading of bacteria on soft agar (swarm) plates on the protein expression level for monocistronic (open symbols, dashed lines) or bicistronic (filled symbols, solid lines) constructs. Protein expression from plasmids harboring *cheB-eyfp* and *cheR_cheB-eyfp* in deletion strain Δ*cheB* and *cheZ-eyfp* and *cheY_cheZ-eyfp* in strain Δ*cheZ* under different IPTG inducer level. The sequence upstream of the *cheY* start codon was modified in the monocistronic construct to achieve expression comparable to the *cheB_cheY-eyfp* construct. Both were expressed in the deletion strain Δ*cheY* under weaker pBAD promoter induction. Expression levels were measured in liquid cultures grown under the same induction as described in the appendix C.1. Chemotaxis efficiency was determined as the size of a swarm rings and normalized to that of wild-type strain. (**b**) Enhancement of chemotactic efficiency by expression coupling was calculated as a ratio of chemotaxis efficiency at a given expression level of the monocistronic construct to the interpolated efficiency at the same expression level of the YFP fusion in the respective bicistronic construct in (**a**), and values at different expression levels were averaged. Error bars indicate standard deviations. Figures with adapted captions are taken from (88).

(fig. 4.5b). We quantified translational coupling by the ratio of the indirect upregulation of the downstream gene to the direct upregulation of the first gene in constructs carrying gene pairs. The determined translational coupling varied from about 0.2 to 0.6, inversely related to the level of translational enhancement of the first gene (fig. 4.5c). This relationship became even more apparent when a stronger *cheR* RBS was used in the *cheR_cheB* pair resulting in a 30 fold translational enhancement and a significantly weaker coupling (~ 0.2) than in the case with an approximately 5-fold enhancement shown in fig. 4.5c.

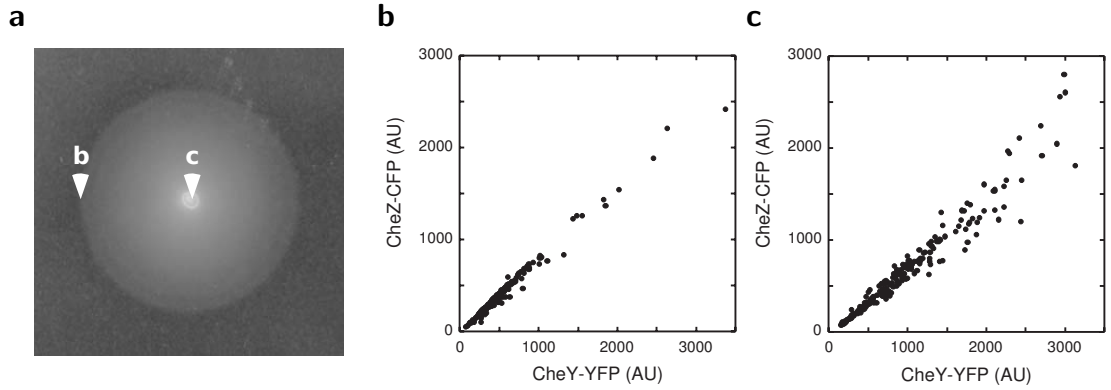


Figure 4.7.: Chemotactic selection for posttranscriptional coupling. (**a**) Chemotaxis-driven spreading of Δ*(cheYcheZ)* cells expressing CheY-YFP and CheZ-CFP from a bicistronic construct on soft agar (swarm) plates. (**b** and **c**) Scatter plots of single-cell levels of CheY-YFP and CheZ-CFP in cells taken from the edge (**b**) and from the middle (**c**) of the spreading colony. Relative concentrations of fluorescent proteins in individual cells were determined using fluorescence microscopy as described in the appendix C.1. Protein expression was induced with 17 μM IPTG. Figures with adapted captions are taken from (88).

4.2.2. Pairwise coexpression of genes improves chemotaxis

Preserving a constant ratio between signaling proteins may be important for the proper functioning of the chemotaxis pathway under fluctuating protein levels. It was previously shown that the chemotaxis system tolerates the concerted overexpression of genes much better than the overexpression of individual genes (89). To test for the effect of overexpression of pairs of genes on the chemotactic performance we coexpressed proteins from bicistronic constructs. We then quantified relative chemotaxis efficiency by measuring the chemotaxis-driven spreading in soft agar (fig. 4.6). Cells that express a YFP fusion to a chemotaxis gene from a monocistronic construct in the corresponding knockout strain showed consistent lower chemotaxis efficiency than cells expressing this fusion as a downstream gene in bicistronic constructs at the same level (fig. 4.6a). Clearly, cells coexpressing pairs of genes spread much more efficiently (fig. 4.6b).

Such an improvement suggest an evolutionary selection for the correlated expression of certain chemotaxis genes. However, the experiment cannot distinguish between transcriptional and translational coupling. Thus we set out to directly test whether there is chemotaxis driven selection for posttranscriptional coupling by monitoring protein levels on the single-cell level. To this end we compared single-cell fluorescent levels of CheY-YFP and CheZ fused to a cyan fluorescent protein, CheZ-CFP. Both were expressed from one bicistronic construct in *E. coli* populations spreading in soft agar (fig. 4.7). Cells best performing chemotaxis at the outer edge of the spreading ring (fig. 4.7a) showed a strong correlation between CheY-YFP and CheZ-CFP expression (fig. 4.7b). On the contrary, non-spreading cells that were not selected for chemotaxis efficiency exhibited a significantly weaker selection

(fig. 4.7c). Since both subpopulations expressed CheY-YFP and CheZ-CFP from the same bicistronic construct, the increased correlation demonstrates chemotactic selection beyond the transcriptional level and supports our assumption that translational coupling should be evolutionary beneficial.

But why are there some proteins coupled through sequential gene arrangement and not others? Why is *cheR* next to *cheB* and not next to *cheZ*? The enhanced correlation of gene pairs is the most likely mechanism by which translational coupling might enhance chemotaxis. We address this question theoretically by simulating the impact of gene order on the robustness of the chemotaxis pathway. To that end we need to simulate the chemotaxis pathway, a detailed model of which is described in the next chapter. Subsequently, a model of translational coupling will be developed which we use to simulate the effect of such coupling on chemotaxis robustness.

4.2.3. Model of the chemotaxis pathway

The chemotaxis pathway of *E. coli* is one of the best investigated bacterial signaling networks. Hence we are in the comfortable situation to know a lot about the protein-protein interactions. We will extend earlier models of the chemotaxis pathway, especially by incorporating the binding of the response regulator CheY and the methylesterase CheB to the kinase CheA.

Receptor model We start with the receptor model developed by Ned Wingreen and coworkers (41, 72). They consider homodimers of two-state receptors, being either active or inactive. The receptor can bind a ligand in the active state

$$\mathrm{T^{on}} + \mathrm{L} \underset{k_-^{\mathrm{on}}}{\overset{k_+^{\mathrm{on}}}{\rightleftharpoons}} [\mathrm{T^{on}L}]$$

or the inactive state

$$\mathrm{T^{off}} + \mathrm{L} \underset{k_-^{\mathrm{off}}}{\overset{k_+^{\mathrm{off}}}{\rightleftharpoons}} [\mathrm{T^{off}L}] \; .$$

Ligands remain bound to the receptor for about one millisecond. The conformation transitions between the active and inactive state of a receptor are thought to be on a microsecond timescale, thus many such transitions occur within a single-ligand binding event (1). The system therefore can be considered to be in thermodynamic equilibrium and the probability to find a receptor in the on-state, given a certain methylation level m and external ligand concentration L, reads

$$p(\mathrm{on}|m;L) = \frac{1}{1 + \left(\frac{1+L/K^{\mathrm{off}}}{1+L/K^{\mathrm{on}}}\right) e^{\epsilon_m}}, \tag{4.1}$$

where $K^{\mathrm{i}} = k^i_-/k^i_+$ denotes the dissociation constant. Attractant binding favors the off-state, thus $K^{\mathrm{off}} < K^{\mathrm{on}}$, whereas energy ϵ_m decreases with methylation m, thus favoring the active state (41, 72) (see appendix C.2 for details).

The receptors are grouped into clusters of N receptors and are assumed to be so tightly coupled that all are either off or on together. Furthermore the cluster contains two types of receptors, N_a Tar- and N_s Tsr-receptors, which differ by their dissociation constants but are assumed to exhibit the same dependence of the free energy on the methylation state. Hence we can group the receptors with the same methylation state and denote their number by n_m. The probability of the whole cluster to be active therefore reads

$$p(\mathrm{on}|n_0, n_1, \ldots, n_k) = \frac{1}{1 + \left(\frac{1+L/K_a^{\mathrm{off}}}{1+L/K_a^{\mathrm{on}}}\right)^{N_a} \left(\frac{1+L/K_s^{\mathrm{off}}}{1+L/K_s^{\mathrm{on}}}\right)^{N_s} \exp\left(\sum_{m=0}^{k} \epsilon_m n_m\right)}. \tag{4.2}$$

This activity is the receptor-clusters output, driving the autophosphorylation of CheA. As the free energy is to good approximation a linear function of the methylation state m, i.e. $\epsilon_m = cm + b$ (see also supplementary figure C.1 in appendix C.2), expression (4.2) can be further simplified

$$F = \sum_{m=0}^{k} \epsilon_m n_m = \sum_{m=0}^{k} n_m(cm + b) = c \sum_{m=0}^{k} n_m m + b \sum_{m=0}^{k} n_m = cM + bN, \tag{4.3}$$

where M is the methylation level of the whole receptor cluster and $N = N_a + N_s$ is the number of receptors. Thus the probability (4.2) to be in an on-state can be written as

$$p_A(M, N) = \frac{1}{1 + \exp(cM + bN)\left(\frac{1+L/K_a^{\mathrm{off}}}{1+L/K_a^{\mathrm{on}}}\right)^{N_a} \left(\frac{1+L/K_s^{\mathrm{off}}}{1+L/K_s^{\mathrm{on}}}\right)^{N_s}}. \tag{4.4}$$

where we introduced $p(\mathrm{on}|M, N) = p_A(M, N)$ for brevity.

Methylation of receptors Methylation of the chemotaxis receptors is catalyzed by CheR, demethylation by phosphorylated CheB (6). Non-phosphorylated CheB is also able to demethylate receptors, but at a much smaller rate and will therefore be neglected (5). Our model makes the assumption that CheR and CheB bind to receptor-clusters independent of their activity states, thus

$$\mathrm{R} + \mathrm{T} \rightleftharpoons [\mathrm{RT}]\,, \tag{4.5}$$

and

$$\mathrm{Bp} + \mathrm{T} \rightleftharpoons [\mathrm{BpT}] \tag{4.6}$$

where T denotes the accessible binding sites of receptor-clusters and R and Bp denote CheR and phosphorylated CheB concentration, correspondingly. For the concentration of CheR and CheB bound to the receptor clusters, which are much more abundant than CheR and

CheB (83), we therefore obtain

$$[RT] = \frac{R^T T}{K^D_{[RT]} + T} \tag{4.7}$$

$$[BpT] = \frac{BpT}{K^D_{[BpT]} + T}, \tag{4.8}$$

where $K^D_{[RT]}$ and $K^D_{[BpT]}$ denote the dissociation constants and $R^T = R + [RT]$ refers to the total concentration of CheR.

In vivo experiments suggest strong localization of chemotaxis proteins (138), hence the dissociation constants in eqs. (4.7) and (4.8) should be much smaller than the receptor concentration, which gives us

$$[RT] \approx R^T \tag{4.9}$$

$$[BpT] \approx Bp. \tag{4.10}$$

Robust adaptation, i.e. independent of the precise fine tuning of kinetic parameters, to a constant external stimulus requires an activity dependent (de)methylation (7, 3). We therefore assume that CheB only demethylates active, whereas CheR mainly methylates inactive receptors. Since receptors are coupled, their activity state only depends on the total methylation level, we do not have to resolve the different receptors within a cluster: The methylation and demethylation of any single receptor within the cluster only changes the total methylation level M of the cluster. Using a mean-field approach, i.e. neglecting the fluctuations of methylation, and assuming that all receptor cluster contain the same number of Tar- and Tsr-receptors $N = N_a + N_s$, we can approximate the mean activity by

$$\langle p_A(M,N)\rangle \approx p_A(\langle M\rangle, N) = \frac{1}{1 + \exp\left(c\langle M\rangle + bN\right)\left(\frac{1+L/K_a^{\mathrm{off}}}{1+L/K_a^{\mathrm{on}}}\right)^{N_a}\left(\frac{1+L/K_s^{\mathrm{off}}}{1+L/K_s^{\mathrm{on}}}\right)^{N_s}}. \tag{4.11}$$

The differential equation for the mean methylation level is then closed

$$\frac{d}{dt}\langle M\rangle = k_R R^T \left[1 - p_A(\langle M\rangle, N)\right](1 - \delta_{\langle M\rangle, M_{\max}}) \tag{4.12}$$

$$- k_{Bp} Bp\, p_A(\langle M\rangle, N)(1 - \delta_{\langle M\rangle,0}), \tag{4.13}$$

where the Kroneckers assure that no (de)methylation above and below the thresholds of $M_{\max}$ and 0 can take place. Solving for the stationary state thus yields the adapted activity which is independent of the ligand concentration and methylation state

$$p_A = \frac{k_R R^T}{k_R R^T + k_{Bp} Bp}. \tag{4.14}$$

Phosphorylation of response regulators The protein kinase CheA consists of several distinct domains, from which two, P1 and P2, are especially important for the phosphorylation of CheY and CheB (148). Both these response regulators bind to the P2 domain. If the P1 domain is phosphorylated, a phosphoryl group can be transferred to bound CheY or CheB (82, 146). Moreover, CheY can be directly phosphorylated without binding to CheA and we assume this to be also valid for CheB (147). In addition, phosphorylated proteins can bind to the P2 domain, but with a lower affinity (82). The scheme of possible reactions for CheY thus reads

$$\begin{array}{ccccc} & & & & \mathrm{A[P_2Yp|P_1p]} & \underset{k_Y^4}{\overset{k_Y^5}{\rightleftharpoons}} & \mathrm{Yp + A[P_2|P_1p]} \\ & & & & k_A p_A \Big\uparrow \Big\downarrow k_Y Y & & k_A p_A \Big\uparrow \Big\downarrow k_Y Y \\ \mathrm{Y + A[P_2|P_1p]} & \underset{k_Y^2}{\overset{k_Y^1}{\rightleftharpoons}} & \mathrm{A[P_2Y|P_1p]} & \xrightarrow{k_Y^3} & \mathrm{A[P_2Yp|P_1]} & \underset{k_Y^4}{\overset{k_Y^5}{\rightleftharpoons}} & \mathrm{Yp + A[P_2|P_1]} \\ k_A p_A \Big\uparrow \Big\downarrow k_Y Y & & k_A p_A \Big\uparrow \Big\downarrow k_Y Y & & & & \\ \mathrm{Y + A[P_2|P_1]} & \underset{k_Y^2}{\overset{k_Y^1}{\rightleftharpoons}} & \mathrm{A[P_2Y|P_1]} & & & & \end{array} \tag{4.15}$$

The reaction scheme for the phosphorylation of CheB looks similar. However, due to its small concentration in *E. coli* (83), we do not take it into account in the conservation relation of CheA, i.e.

$$\begin{aligned} A^T &= A[P_2|P_1] + A[P_2|P_1p] + A[YP_2|P_1] + A[YpP_2|P_1p] + A[BP_2|P_1] + A[BpP_2|P_1p] \\ &\approx A[P_2|P_1] + A[P_2|P_1p] + A[YP_2|P_1] + A[YpP_2|P_1p]. \end{aligned} \tag{4.16}$$

Hence terms involving the phosphorylation of CheB will not show up in the rate equations for CheA. It is also not necessary to consider bimolecular reactions of CheY with CheA, given CheB bound to the P2 domain, in contrast to the bimolecular reactions of CheB with CheA, where terms of the form $[B]A[Y \cdot |P_1p]$ are important.

Phosphorylation of the P1 domain is driven by the activity of receptor clusters. We therefore model the rate of CheA phosphorylation by

$$k_A p_A A[P_2 \cdot |P_1]. \tag{4.17}$$

Dephosphorylation of CheY is mediated by the phosphatase CheZ in a Michaelis Menten reaction (14)

$$\mathrm{Yp + Z} \underset{k_Z^2}{\overset{k_Z^1}{\rightleftharpoons}} \mathrm{[YpZ]} \xrightarrow{k_Z^3} \mathrm{Y + Z + p}. \tag{4.18}$$

For CheB no phosphatase could be found, and thus it is believed that CheB dephosphorylates spontaneously (Sourjik, personal communication). The full set of equations and applied approximations can be found in the appendix C.2. In order to asses robustness of the

pathway these equations were solved for the stationary state of phosphorylated CheY.

4.2.4. Modeling translational coupling

Now that we have established a model of the chemotaxis pathway, we need to develop a framework to account for translational coupling between chemotaxis genes. This will then be used to rank chemotaxis performance of different permutations of the chemotaxis genes *cheR, cheB, cheY* and *cheZ* from the *meche* operon. The point of departure is an extension of the general framework to describe the translation process as outlined in the background chapter 2.

Modeling of protein-fluctuations and correlations

We now consider a polycistronic mRNA coding for S different proteins. We model accessibility A_i of the ith ribosome binding site by a random telegraph process, i.e. stochastic switching of the RBS between only two states, an accessible (open) $\mathrm{m_i^o}$ and an inaccessible (closed) state $\mathrm{m_i^c}$,

$$\mathrm{m_i^c} \underset{k_i^c}{\overset{k_i^o}{\rightleftharpoons}} \mathrm{m_i^o} \ . \tag{4.19}$$

In accordance with the findings described in chapter 3 we assume that accessibility is mainly determined by the folding of the mRNA in the region of the ribosome binding site. The dynamics are captured by the differential equations

$$\frac{d}{dt} m_i^c = -k_i^o m_i^c + k_i^c m_i^o \tag{4.20}$$

$$\frac{d}{dt} m_i^o = +k_i^o m_i^c - k_i^c m_i^o. \tag{4.21}$$

Since this process can be assumed to be very fast, a rapid equilibrium is achieved and we can solve for the stationary state

$$m_i^c = m_i^f \frac{k_i^c}{k_i^c + k_i^o} = m_i^f \langle 1 - A_i \rangle \tag{4.22}$$

$$m_i^o = m_i^f \frac{k_i^o}{k_i^c + k_i^o} = m_i^f \langle A_i \rangle, \tag{4.23}$$

where we used the quasi-steady state conservation relation for the concentration of free ribosome binding sites, m_i^f

$$m_i^f = m_i^c + m_i^o \tag{4.24}$$

which holds under the assumption that folding of the RBS occurs on the fastest timescale in the system. If we now assume that ribosomes only bind to the accessible RBS, we have

to modify our scheme (2.22)

$$\mathrm{m_i^o} + \mathrm{R} \underset{k_i^-}{\overset{k_i^+}{\rightleftharpoons}} [\mathrm{R_{in}}|\mathrm{m_i}] \xrightarrow{k_i^{cat}} [\mathrm{R_{el}}|\mathrm{m_i}] + \mathrm{m_i^f} \xrightarrow{v_i} \mathrm{p} + \mathrm{R} + \mathrm{m_i^f}\ . \tag{4.25}$$

Due to eq. (4.23) we have to modify eqs. (2.27) – (2.28), now reading

$$(m_i - [R_{\mathrm{in}}|m_i])\langle A_i\rangle R - K_i^M [R_{\mathrm{in}}|m_i] \approx 0 \tag{4.26}$$

$$k_i^{cat}[R_{\mathrm{in}}|m_i] - v_i[R_{\mathrm{el}}|m_i] \approx 0. \tag{4.27}$$

Solving for $[R_{\mathrm{in}}|m_i]$ and $[R_{\mathrm{el}}|m_i]$ we obtain

$$[R_{\mathrm{in}}|m_i] = \frac{m_i\langle A_i\rangle R}{K_i^M + \langle A_i\rangle R} \tag{4.28}$$

$$[R_{\mathrm{el}}|m_i] = \frac{k_i^{cat}}{v_i}\frac{m_i\langle A_i\rangle R}{K_i^M + \langle A_i\rangle R}. \tag{4.29}$$

For the sake of simplicity the average number of ribosomes elongating on the ith ORF of a single transcript will be denoted by $\langle n_i\rangle$, which is related to $[R_{\mathrm{el}}|m_i]$ by

$$\langle n_i\rangle = \frac{[R_{\mathrm{el}}|m_i]}{m_i} = \frac{k_i^{cat}}{v_i}\frac{\langle A_i\rangle R}{K_i^M + \langle A_i\rangle R}. \tag{4.30}$$

Since

$$v_i = \frac{1}{\sum_{j=1}^{L_i} v_{ij}^{-1}} \sim \frac{v}{L_i}, \tag{4.31}$$

where v is the mean elongation rate per codon, we obtain for the average fraction $\langle x_i\rangle$ of codons being translated by the elongating ribosomes

$$\langle x_i\rangle = \frac{\langle n_i\rangle}{L_i} \sim \frac{k_i^{cat}}{v}\frac{\langle A_i\rangle R}{K_i^M + \langle A_i\rangle R}. \tag{4.32}$$

Note that $\langle x_i\rangle$ is independent of the number of codons. Translational coupling arises by the dependence of the accessibility $A_i = A_i(t, x_{i-1})$ on the upstream ribosome density x_{i-1}. The rationale is that the higher the ribosome density on the upstream gene the more ribosomes start translation of the downstream gene as a consequence of an increase in $A_i(t, x_{i-1})$. This simple model is able to account for the experimental fact that an upstream gene with weak ribosome binding site can significantly influence translational efficiency of a downstream gene with strong ribosome binding site.

The assumption that translational coupling works predominantly in downstream direction leads to a change in translational rate of the ith gene, Δx_i, in response to an externally induced increase in ribosome density, $\Delta x_{i-1}^{\mathrm{ext}}$, of the upstream gene, e.g. caused by mutants

with stronger ribosome binding sites. To linear order this response is given by

$$\frac{\Delta x_i}{\langle x_i \rangle} = \alpha_{ii-1} \frac{\Delta x_{i-1}^{\text{ext}}}{\langle x_{i-1} \rangle} \qquad \text{with} \quad \alpha_{ii-1} < 1 \tag{4.33}$$

with $\langle x_i \rangle$ the ensemble average of the unperturbed ribosome density. The coupling constants depend on changes relative to the average ribosome density but not on their absolute values, i.e.

$$\alpha_{ii-1} = \frac{d \log(\langle x_i \rangle)}{d \log(\langle x_{i-1} \rangle)} \approx \frac{\text{te}^{\text{ind}} - 1}{\text{te}^{\text{dir}} - 1}, \tag{4.34}$$

where te^{dir} and te^{ind} denote the direct and indirect translational enhancement.

For the downstream RBS it makes no difference whether an increase in the ribosome density is due to a systematic deviation or due to a fluctuation in the initiation at the upstream gene. Hence we expect this linear response relation to hold also for stochastic fluctuations in the ribosome density such that Δx_i^{ext} can be substituted by the intrinsic noise contribution of the corresponding gene. The general coupling between any two genes is given by

$$\Delta x_i = A_{ij} \Delta x_j. \tag{4.35}$$

This means that the fluctuations in ribosome density of each gene are modeled as a superposition of intrinsic fluctuations and contributions from upstream genes via translational coupling. Here we assume equal response coefficients $\alpha_{i,i-1} \approx \alpha$. In this case the structure of the response matrix $\mathbf{A}$ reads

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & \dots & & & 0 \\ \alpha \frac{\langle x_2 \rangle}{\langle x_1 \rangle} & 1 & 0 & \dots & & 0 \\ \alpha^2 \frac{\langle x_3 \rangle}{\langle x_1 \rangle} & \alpha \frac{\langle x_3 \rangle}{\langle x_2 \rangle} & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ \alpha^{S-1} \frac{\langle x_S \rangle}{\langle x_1 \rangle} & \alpha^{S-2} \frac{\langle x_S \rangle}{\langle x_2 \rangle} & \alpha^{S-3} \frac{\langle x_S \rangle}{\langle x_3 \rangle} & \alpha^{S-4} \frac{\langle x_S \rangle}{\langle x_4 \rangle} & \dots & 1 \end{pmatrix}. \tag{4.36}$$

The matrix reflects that relative changes in ribosome density ℓ genes upstream of a given gene contribute to the relative changes in translational efficiency attenuated by a factor α^ℓ. As the response coefficient is significantly smaller than one ($\alpha \approx 0.25$) effects on translational efficiency arise predominantly from adjacent genes (see also figs. 4.5b and 4.5c).

Fluctuations in ribosome density due to stochastic binding

In the following we estimate the fluctuations in the stationary ribosome densities x_i originating from stochastic independent binding events of ribosomes. The stationary probability

for the number n_i of ribosomes on the gene i is in this case given by a binomial distribution

$$p(n_i) = \binom{L_i}{n_i} (\langle x_i \rangle)^{n_i} (1 - \langle x_i \rangle)^{L_i - n_i} . \tag{4.37}$$

As the ribosome density is in general small, $\langle x_i \rangle \ll 1$ and $\langle n_i \rangle = \langle x_i \rangle L_i \ll L_i$, n_i follows a Poisson distribution to good approximation

$$p(n_i) \approx \frac{\langle n_i \rangle^{n_i}}{n_i!} e^{-\langle n_i \rangle} \tag{4.38}$$

that in turn can approximated by a Gaussian distribution for $\langle n_i \rangle > 10$

$$p(x_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(x_i - \langle x_i \rangle)^2}{2\sigma_i^2} \right], \quad \sigma_i^2 = \frac{\langle x_i \rangle}{L_i}. \tag{4.39}$$

By eq. (4.35) the ribosome densities x_i depend on the upstream ribosome densities $x_i = x_i(x_{i-1})$. Translational coupling thus leads to a joint probability distribution of the form

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{S}{2}} \sqrt{\det(\boldsymbol{C})}} \exp\left[-\frac{1}{2} \Delta \mathbf{x}^T \cdot \boldsymbol{C}^{-1} \cdot \Delta \mathbf{x} \right], \tag{4.40}$$

with the covariance matrix

$$\boldsymbol{C} = \mathbf{A} \cdot \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_S^2) \cdot \mathbf{A}^T. \tag{4.41}$$

Consequences for protein copy numbers

On the translation time scale the dynamics of the protein copy number $P_i(t)$ can be captured by the simple differential equation

$$\frac{d}{dt} P_i(t) = v \left(\langle x_i \rangle + \Delta x_i(t) \right) - \gamma_p P_i(t). \tag{4.42}$$

Here v denotes the average elongation speed and γ_p the dilution rate given by $\gamma_p = \ln(2)/\tau$, with τ being the generation time (see also appendix A.1). We separate the mean and fluctuations using $x_i = \langle x_i \rangle + \Delta x_i(t)$, with $\langle \Delta x_i(t) \rangle = 0$. We assume further that the correlation time $1/\beta$ is about the same for all genes, i.e. $\langle \Delta x_i(t) \Delta x_j(t') \rangle = C_{ij} e^{-\beta |t - t'|}$ for all i, j. The solution of eq. (4.42) yields

$$P_i(t) = \int_{-\infty}^{t} e^{-\gamma_p (t - t')} v \left(\langle x_i \rangle + \Delta x_i(t') \right) dt'. \tag{4.43}$$

Since integration is a linear operation the distribution for P_i is also Gaussian. The expectation value of P_i corresponds to the population average and is given by

$$\langle P_i(t)\rangle = \int_{-\infty}^{t} e^{-\gamma_p(t-t')} v\langle x_i\rangle dt' = \frac{v}{\gamma_p}\langle x_i\rangle. \tag{4.44}$$

Due to eq. (4.32), the number of proteins being expressed from one mRNA only depends on the strength of its ribosome binding site and the translation initiation rate but not on the elongation speed. The correlation between the ith and jth gene product, P_i and P_j, is given by

$$\begin{aligned}
\langle P_i(t)P_j(t)\rangle &= \left\langle \int_{-\infty}^{t} e^{-\gamma_p(t-t')} v\left(\langle x_i\rangle + \Delta x_i(t')\right) dt' \int_{-\infty}^{t} e^{-\gamma_p(t-t'')} v\left(\langle x_j\rangle + \Delta x_j(t'')\right) dt'' \right\rangle \\
&= \left\langle \int_{-\infty}^{t} dt' \int_{-\infty}^{t} dt'' e^{-\gamma_p(2t-t'-t'')} v^2 [\langle x_i\rangle\langle x_j\rangle + \langle x_j\rangle \Delta x_i(t') \right. \\
&\qquad \left. + \langle x_i\rangle \Delta x_j(t'') + \Delta x_i(t')\Delta x_j(t'')] \right\rangle \\
\langle P_i(t)P_j(t)\rangle &= \langle P_i(t)\rangle\langle P_j(t)\rangle + v^2 \int_{-\infty}^{t} dt' \int_{-\infty}^{t} dt'' e^{-\gamma_p(2t-t'-t'')} \langle \Delta x_i(t')\Delta x_j(t'')\rangle \\
&= \langle P_i(t)\rangle\langle P_j(t)\rangle + v^2 C_{ij} \int_{-\infty}^{t} dt' \int_{-\infty}^{t} dt'' e^{-\gamma_p(2t-t'-t'')} e^{-\beta|t'-t''|} \\
&= \langle P_i(t)\rangle\langle P_j(t)\rangle + v^2 C_{ij} \int_{-\infty}^{t} dt' \left[\frac{e^{-2\gamma_p(t-t')}}{\beta+\gamma_p} + \frac{e^{-2\gamma_p(t-t')}}{\beta-\gamma_p} - \frac{e^{-(\beta+\gamma_p)(t-t')}}{\beta-\gamma_p} \right] \\
&= \langle P_i(t)\rangle\langle P_j(t)\rangle + \frac{v^2 C_{ij}}{\gamma_p(\beta+\gamma_p)}.
\end{aligned} \tag{4.45}$$

Since we assumed equal correlation times, the covariance matrix is given by

$$\Xi_{ij} = \langle P_i(t)P_j(t)\rangle - \langle P_i(t)\rangle\langle P_j(t)\rangle = \frac{v^2}{\gamma_p(\gamma_p+\beta)} C_{ij}. \tag{4.46}$$

The stationary probability distribution for the copy numbers P_i finally reads

$$p(\mathbf{P}) = \frac{1}{(2\pi)^{\frac{S}{2}}\sqrt{\det(\boldsymbol{\Xi})}} \exp\left[-\frac{1}{2} (\mathbf{P} - \langle \mathbf{P}\rangle)^T \cdot \boldsymbol{\Xi}^{-1} \cdot (\mathbf{P} - \langle \mathbf{P}\rangle) \right]. \tag{4.47}$$

Intrinsic noise due to stochasticity in translation

The timescale for fluctuations in ribosome density on gene i is of the order the time it takes to translate a protein, thus $\beta \approx v/L \approx 0.1\,\mathrm{s}^{-1}$, where we assumed a translational speed of $v = 40\,\mathrm{aa}$ per second (see table 2.1 in chapter 2) and a typical gene length of $L = 400\,\mathrm{aa}$.

The variance of fluctuations in protein copy number can be calculated by using (4.46) and the approximation $\frac{1}{\gamma_p(\beta+\gamma_p)} \approx \frac{1}{\gamma_p\beta}$. Hence, for the case without translational coupling we

obtain for the variance of expression-level i

$$\sigma_{P_i}^2 = \frac{v}{\beta}\frac{\langle P_i \rangle}{L_i} \approx \langle P_i \rangle \tag{4.48}$$

and for the relative fluctuations

$$\eta_i = \frac{\sigma_{P_i}}{\langle P_i \rangle} \approx \frac{1}{\sqrt{\langle P_i \rangle}}. \tag{4.49}$$

So far we have considered the situation where only one mRNA is present in the cell at every instant of time. If the same amount of protein, $\langle P_i \rangle$, is synthesized by m mRNAs the noise in proteins synthesized exclusively from the k-th mRNA is given by $\sigma_{P_ik}^2 = \langle P_i \rangle / m$. However, the variance of the total protein copy number in the cell is given again by $\sigma_{P_i}^2 = \sum_{k=1}^{m} \sigma_{P_ik}^2 = \langle P_i \rangle$ because of stochastic independence of the translational events on each mRNA. The variations in protein copy number of CheY and CheZ from a bicistronic plasmid should reflect the translational noise as proteins are diluted by cell division. However, the measured standard deviation over mean in protein copy number of ≈ 0.2 is far beyond the expected value from our analysis $\sigma_{P_Z}/\langle P_Z \rangle \approx 1/\sqrt{2500} = 0.02$ (73). Hence our assumption of a binomially distributed (eq. (4.37)) ribosome occupancy underestimates the variance. However, as long as we can approximate fluctuations of the ribosome density by normally distributed variables, treating the variance as an effective parameter, the derived relations eqs. (4.44) and (4.46) and correspondingly eq. (4.47) continue to hold. Thus we will use eq. (4.47) for modeling intrinsic translational noise and coupling, but take into account a larger variance.

Actually we are interested in the protein concentration c_i, hence we have to divide the copy number P_i by the cell volume V_c,

$$c_i = \frac{P_i}{V_c}. \tag{4.50}$$

In addition to the intrinsic noise arising from translation, the pathway is also exposed to extrinsic noise, which may arise from random variations in RNA polymerase or ribosome abundance. Due to the transcriptional coupling, also fluctuations in the mRNA level are extrinsic to the system. These noise sources cause correlated fluctuations in the protein concentrations. However, as we will see later, these correlated variations should not impact the ranking of gene orders. Moreover, the pathway was shown to be robust against the concerted overexpression of genes (73). Hence it is sufficient to take into account only intrinsic noise and protein concentrations are thus generated according to

$$\mathbf{c} = \nu\, \mathbf{A} \cdot \mathrm{diag}\left(\langle c_1 \rangle \theta_1, \langle c_2 \rangle \theta_2, \ldots, \langle c_S \rangle \theta_S\right) \cdot \boldsymbol{\xi}^{in}, \tag{4.51}$$

with

$$\eta_i = \nu\theta_i = \frac{\sigma_{P_i}}{\langle P_i \rangle}, \quad \langle c_i \rangle = \frac{\langle P_i \rangle}{V_c}. \tag{4.52}$$

The parameter ν controls the amount of intrinsic noise and θ_i specifies the differences in relative noise. The relative noise levels $\nu\theta_i$ are parameters and are not calculated from the actual protein copy-numbers. Most of the time we approximate $\theta_i = 1$. In order to determine the response matrix $\mathbf{A}$, eqs. (4.44) and (4.52) were used to substitute $\langle x_i \rangle$ in eq. (4.36).

4.2.5. Translational coupling between selected genes is predicted to enhance robustness of the pathway

Now that we have developed a model for translational coupling and the chemotaxis pathway, we can move on to the question whether selection for chemotaxis robustness has an influence on the ordering of genes within the *meche* operon.

As noted in the introduction, *E. coli* performs chemotaxis by a biased random walk. Swimming runs are interrupted by tumbling events, i.e. events that partially randomize the direction the bacterium is moving. By tuning the frequency of tumbling events in response to changing attractant concentrations the bacteria swims into favorable directions and avoids swimming into unfavorable ones. Swimming is accomplished by a concerted rotation of the flagellar motors in the counter-clockwise sense (CCW), whereas clockwise (CW) rotation of the motors leads to tumbling. The CW bias is controlled by the concentration of free phosphorylated CheY and follows a steep response curve with a Hill coefficient of about 10 (24). As argued in the introduction, maintaining a certain CW bias in the adapted state is a reliable measure for chemotactic performance.

In our setup, we solve the stationary state equations (see section 4.2.3 and appendix C.2) governing the chemotaxis pathway subject to different total concentrations of the involved proteins as determined by stochastic gene expression (4.51). The cell-to-cell variations of the protein concentrations lead to a distribution in the adapted CheYp level within a population and hence to a distribution in the CW bias, calculated as

$$\text{CW-bias} = \frac{(Yp)^{10}}{(K_h)^{10} + (Yp)^{10}}. \tag{4.53}$$

Within the *meche* operon, all 24 permutations of the gene order of the chemotaxis proteins CheR, CheB, CheY and CheZ are simulated according to eq. (4.51). For each sample consisting of 10^5 cells the value of K_h was determined such that the physiological value of the average CW bias is $\langle \text{CW-bias} \rangle = 0.2$. The standard deviation in CW bias is a measure of the chemotaxis efficiency within a population. We therefore rank the different gene permutations according to the standard deviation in CW bias.

Additionally, we performed simulations with proteins that are perfectly coupled in gene expression. Perfect coupling between concentrations of any two proteins, c_1 and c_2, results

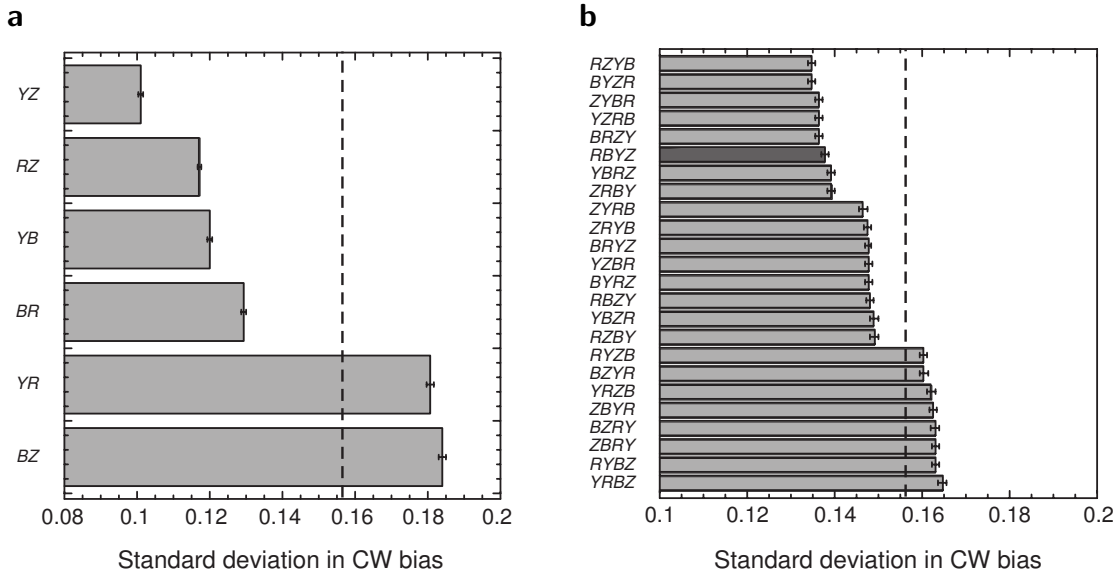


Figure 4.8.: Simulated effects of translational coupling on robustness of the signaling output. Standard deviation of the CW motor bias in a population of 10^5 cells was simulated in presence of gene expression noise. (**a**) Simulations for 100% pairwise coupling of indicated chemotaxis genes, with remaining genes being uncoupled. (**b**) Simulations for different arrangements of translationally coupled chemotaxis genes, performed at equal noise levels for all genes and 25% coupling. Genes are indicated by single letters, i.e., Y = CheY, and so forth. Error bars indicate confidence intervals obtained by bootstrapping. Figures with adapted captions are taken from (88).

from the relation

$$c_2 = \langle c_2 \rangle \times \frac{c_1}{\langle c_1 \rangle}, \tag{4.54}$$

with c_1 a stochastic concentration in an otherwise uncoupled randomly fluctuating background. The stochastic variables were taken from a multi-Gaussian distribution (4.51) with $\mathbf{A} = \mathbf{1}$ being the unit matrix, $\nu = 0.05$ and $\theta_i = 1$.

The results are shown in fig. 4.8a. The dotted line in this bar-plot shows the standard deviation in CW bias without any coupling. Most of the pairings lead to a decrease of the standard deviation, but interestingly two of them lead to worse chemotactic performance. We can conclude that there are favorable and unfavorable pairings of chemotaxis genes depending on whether they decrease or increase the standard deviation in CW bias.

Taking this into account we can understand the ranking of the 24 permutations shown in fig. 4.8b. The noise was generated by using a coupling parameter $\alpha = 0.25$, $\nu = 0.05$ and $\theta_i = 1$. Three blocks of permutation can be identified (fig. 4.8b). They differ in respect to the number of favorable pairings. The first block reaching from *RZYB* to *ZRBY* consists of permutation with no unfavorable pairings. The second and third block, reaching from *ZYRB* to *RZBY* and *RYZB* to *YRBZ*, contain permutations with one and two unfavorable

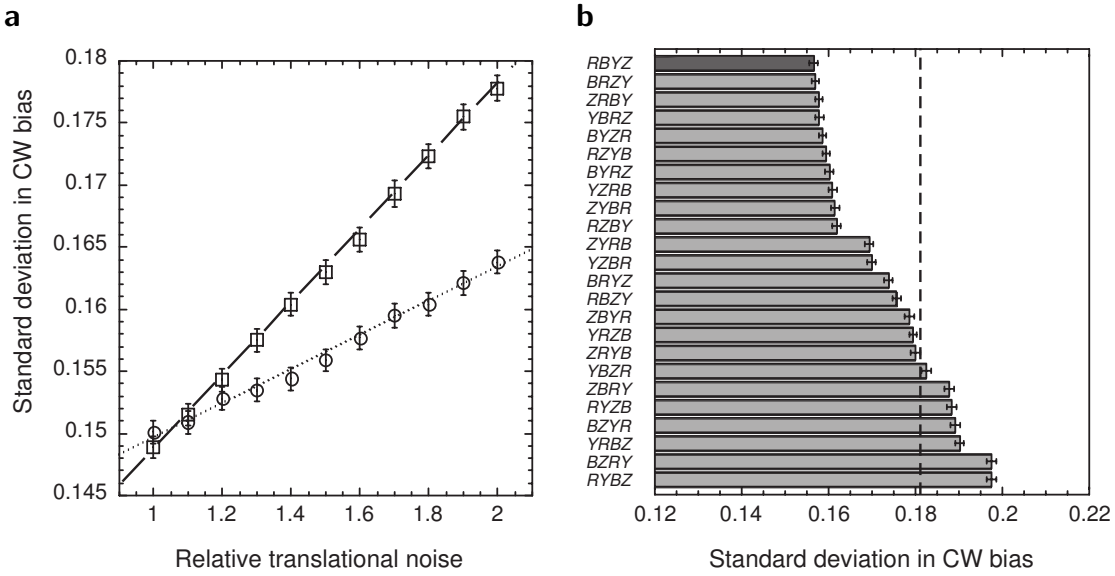


Figure 4.9.: Effect of different relative noise levels on chemotactic performance. (**a**) Asymmetric effects of translational noise for 25% coupling between *cheR_cheZ* (circles, dotted line) and *cheZ_cheR* (squares, dashed line) observed for different noise levels of CheR with the remaining genes being uncoupled. Linear fits to the data are guide to the eye. (**b**) Simulations for different gene orders as in fig. 4.8b, at 1.5-fold higher noise for the weakly expressed *cheR* and *cheB* genes. Dark-gray bars indicate gene order in *E. coli*. Standard deviation of CW bias in absence of coupling is indicated by vertical dashed lines. Error bars indicate confidence intervals obtained by bootstrapping. Figures with adapted captions are taken from (88).

pairings, respectively. Apart from this gene order does not matter in fig. 4.8b, as can be seen by the fact that each permutation and its counterpart arising from reflection of the gene order perform equally well within statistical fluctuations. Explanation of the ranking order on grounds of favorable an unfavorable pairings is possible, as in our model correlations in expression levels decay exponentially fast

$$\varrho(c_i, c_j) \sim \alpha^{|i-j|} + \mathcal{O}(\alpha^{|i-j|+2}). \tag{4.55}$$

However, the sequence of genes in pairs becomes important if noise levels are different (fig. 4.9a). Thus the ranking order changes if we assume a higher relative noise for CheR and CheB, i.e. $\theta_R = \theta_B = 1.5$ (fig. 4.9b). In this case order of pairings matters, as for example *BYRZ* and *ZRYB*, previously ranked equally, now occupy different places. The permutation *BYRZ* is ranked much higher compared to *ZRYB*. The reason for this is of course the difference in noise levels, leading to distinct correlation coefficients depending on the order of pairings

$$\varrho(c_{i-1}, c_i) = \alpha \frac{\theta_{i-1}}{\theta_i} + \mathcal{O}(\alpha^3). \tag{4.56}$$

This can intuitively be understood if one considers the case where one of the proteins is expressed with no noise at all. First, assume the noisy protein in the first place. Due to the coupling, the non-fluctuating one is able to follow the translational fluctuations of the first protein closely, however this is not the case if we reverse the order. In this case the second protein is still fluctuating and therefore correlations are smaller compared to the first case.

From the expression (4.56) we see that depending on the correlation coefficient pairings are weighted differently. Hence the effect of beneficial pairings can be amplified, whereas the effect of unfavorable pairings can be attenuated. Consider our example from above, i.e *BYRZ* and *ZRYB*. For the first permutation, the correlation coefficients calculated by (4.56) with $\theta_R = \theta_B = 1.5$ and $\theta_Y = \theta_Z = 1$ read

$$\varrho(\mathrm{B}, \mathrm{Y}) \sim 0.375 \tag{4.57}$$

$$\varrho(\mathrm{Y}, \mathrm{R}) \sim 0.167 \tag{4.58}$$

$$\varrho(\mathrm{R}, \mathrm{Z}) \sim 0.375. \tag{4.59}$$

In contrast, the correlation coefficients for *ZRYB* are

$$\varrho(\mathrm{Z}, \mathrm{R}) \sim 0.167 \tag{4.60}$$

$$\varrho(\mathrm{R}, \mathrm{Y}) \sim 0.375 \tag{4.61}$$

$$\varrho(\mathrm{Y}, \mathrm{B}) \sim 0.167. \tag{4.62}$$

Comparing these numbers, we see that in the permutation *BYRZ* the favorable pairings *BY* and *RZ* are weighted stronger than the unfavorable pairing *YR*. For the permutation *ZRYB* it is exactly the other way round, thus it is ranked much lower.

In addition to the gene order effect, the simple pattern of blocks formed by the permutations with a given number of unfavorable pairings has vanished, the permutation *BYRZ* is enclosed by *RZYB* and *YZRB*. The two latter ones comprise only favorable pairings in contrast to the former one which contains the unfavorable pair *YR*. Taking the other relevant correlation coefficients into account

$$\varrho(\mathrm{Z}, \mathrm{Y}) \sim \varrho(\mathrm{Y}, \mathrm{Z}) \sim \varrho(\mathrm{R}, \mathrm{B}) \sim 0.25. \tag{4.63}$$

we learn that only one of the pairings, *RZ*, is weighted by a higher correlation coefficient of ~ 0.375, whereas the other pairings have a correlation coefficient of ~ 0.25 or even ~ 0.167. Hence the larger correlation of the beneficial pairings and the attenuation of the bad pair in the case of *BYRZ* seems to be sufficient to make it equivalent to permutations with only good pairings.

In summary we can conclude that if there are differences in noise level between the proteins, the ranking of gene order is altered by the relation (4.56). Attenuation of disadvantageous

and amplification of beneficial pairings have to be taken into account to understand the effects on chemotactical performance.

An additional effect which could be important for the determination of the gene-order is the propagation of noise. Due to translational coupling the standard deviations of proteins at the backmost positions within the operon become larger. In our model, however, this effect is only of second order in α.

Since the CW bias is determined by the free Yp level within the cell, we will do a linear approximation of noise propagation for the free Yp concentration with respect to fluctuations of the total concentrations of CheA, CheY, CheZ, CheR and CheB. If the fluctuations, denoted by $\delta A^T, \delta Y^T, \delta Z^T, \delta R^T$ and δB^T, are sufficiently small, it is possible to determine the ranking of permutations by this expansion.

The expansion of Yp up to linear order is given by

$$\delta Yp = \frac{\partial Yp}{\partial A^T}\delta A^T + \frac{\partial Yp}{\partial Y^T}\delta Y^T + \frac{\partial Yp}{\partial Z^T}\delta Z^T + \frac{\partial Yp}{\partial R^T}\delta R^T + \frac{\partial Yp}{\partial B^T}\delta B^T. \tag{4.64}$$

Calculating the variance of Yp yields

$$\langle\langle Yp^2\rangle\rangle = \langle(\delta Yp)^2\rangle. \tag{4.65}$$

We numerate the proteins according to their position on the operon and use c_i to indicate the total concentration of the ith protein (compare section 4.2.4). Furthermore, CheA concentration is denoted by c_0 and there is no translational coupling between the kinase and the remaining chemotaxis proteins. Thus, we get for the variance up to linear order in the fluctuating proteins

$$\langle(\delta Yp)^2\rangle = \sum_{i=0}^{4}\left(\frac{\partial Yp}{\partial c_i}\right)^2\langle(\delta c_i)^2\rangle + \sum_{\substack{i,j=1\\ i\neq j}}^{4}\frac{\partial Yp}{\partial c_i}\frac{\partial Yp}{\partial c_j}\langle\delta c_i\delta c_j\rangle. \tag{4.66}$$

All derivatives are performed at the average wild-type protein concentration. The first sum refers to the noise in the system and the second to the correlations. We can now separate the influences on the variance of Yp. The derivatives give the response of the biochemical network to the perturbations in protein concentrations, whereas $\langle(\delta c_i)^2\rangle$ and $\langle\delta c_i\delta c_j\rangle$ reflect the statistical properties of the gene expression process.

Using our model for gene expression noise (4.51) and neglecting terms in quadratic or higher order in the coupling parameter α, we only get correlation between next neighbors,

$$\langle(\delta Yp)^2\rangle = \sum_{i=0}^{4}\left(\frac{\partial Yp}{\partial c_i}\right)^2\langle(\delta c_i)^2\rangle + 2\sum_{i=2}^{4}\frac{\partial Yp}{\partial c_{i-1}}\frac{\partial Yp}{\partial c_i}\langle\delta c_{i-1}\delta c_i\rangle + \mathcal{O}(\alpha^2), \tag{4.67}$$

and the averages in this approximation read

$$\langle(\delta c_i)^2\rangle = \nu^2\theta_i^2\langle c_i\rangle^2 + \mathcal{O}(\alpha^2) \tag{4.68}$$

$$\langle\delta c_{i-1}\delta c_i\rangle = \alpha\nu^2\theta_{i-1}^2\langle c_{i-1}\rangle\langle c_i\rangle + \mathcal{O}(\alpha^3). \tag{4.69}$$

Here, we assumed that there is no extrinsic noise in the system, as taking it into account does not change the ranking of permutations, i.e. the effects of translational coupling. Extrinsic noise just adds to each of the normalized covariances $covar(c_i, c_j)/(\langle c_i\rangle\langle c_j\rangle)$ the variance of the extrinsic noise and therefore does not lead to a differential ranking.

This approximation is sufficient to determine the ranking of the different permutations. Using $\sqrt{\langle(\delta Yp)^2\rangle}$ as given by eq. (4.67), we recover essentially the same results as from the simulation. The derivatives were calculated numerically by using the model introduced in section 4.2.3.

Depending on the sign of the product $\frac{\partial Yp}{\partial c_{i-1}}\frac{\partial Yp}{\partial c_i}$, the correlations lead to an in- or decrease of the standard deviation of Yp, since the quadratic terms are always larger than zero. Independent of model details we can assume

$$\frac{\partial Yp}{\partial Y^T} > 0, \qquad \frac{\partial Yp}{\partial Z^T} < 0 \tag{4.70}$$

$$\frac{\partial Yp}{\partial R^T} > 0, \qquad \frac{\partial Yp}{\partial B^T} < 0. \tag{4.71}$$

Increasing the total CheY concentration means there is more CheY that can be phosphorylated. On the contrary more phosphatase CheZ in the system implies a decreasing Yp level. The enzymes CheR and CheB regulate the methylation of receptors and therefore the steady state receptor-activity, which affects the level of phosphorylated CheY. Increasing the CheR level leads to an increase in kinase activity, whereas increasing CheB concentration reduces kinase activity. Hence we concluded that

$$YZ \quad RB \quad YB \quad ZR \tag{4.72}$$

are favorable pairings, since the corresponding products of derivatives are negative. In contrast

$$YR \quad ZB \tag{4.73}$$

are unfavorable, as the product of derivatives is positive. Thus we recover the result from the analysis of pairwise coupling.

If all proteins show the same noise-level, i.e. θ_i is the same for all, the order of pairs does not matter. However, with different noise level the correlation is larger if the protein fluctuating stronger is in the first place, i.e. $\theta_{i-1} > \theta_i$, since only θ_{i-1} shows up in expression (4.69) for the correlation. Therefore the impact of favorable pairings can be amplified and those of unfavorable ones, both compared to the overall noise level, can be attenuated.

4.3. Discussion

4.3.1. Translational coupling as a mechanism of noise reduction

The fluctuation of protein levels in genetically homogeneous cell populations, referred to as gene expression noise, is one of the most important perturbations affecting the efficacy of cellular pathways. It arises from the variation of global factors, e.g. the number of ribosomes, influencing the expression of all genes and the stochastic nature of promoter activity (110, 116, 117, 121). Due to the organization of genes coding for proteins of related functions in operons and the control of transcription by common regulators, expression of such genes is coupled on the transcriptional level in bacteria. As a result of this coupling we expect the concerted variation of related genes to be the dominant form of gene expression noise in bacteria. Indeed a strong correlation was observed between proteins of the chemotaxis system in *E. coli* and the pathway topology was shown to be particularly robust against such coupled fluctuations in its components (73). However, the stochasticity of translation, most probable due to the fluctuations in translation initiation, gives rise to uncorrelated variations of the protein levels even if their genes are encoded in a single transcriptional unit (73). Since pathway topologies may have evolved to compensate primarily correlated variations of protein levels, such uncorrelated noise might adversely affect their functioning. It is therefore to be expected that bacteria evolved additional means to mitigate the harmful effects of translational noise. Translational coupling is such a mechanism. It was described before, especially in metabolic operons (109, 126, 119, 48, 86), but also between genes encoding ribosomal proteins (8) and a two component sensor (85). Coupling most likely happens when the termination codon of the upstream gene is located near the start codon or the Shine-Dalgarno (SD) sequence of the downstream gene. However, it might also take place due to long-range RNA interactions regulating translation via feedback mechanisms (21, 22, 157). Several factors may contribute to translation coupling. First, the translation of the upstream will lead to a local increase of the number of ribosomes near the start codon of the downstream gene. This in turn could allow for a more efficient reinitiation of translation even if a strong SD sequence is lacking (48). Second, ribosomes translating the upstream gene may unfold local RNA structures burying the SD sequence or the start codon of the downstream gene. Increasing the accessibility of the RBS can facilitate both reinitiation of ribosomes already translating the upstream gene or de-novo binding of ribosomes to the transcript (119). The latter model is supported by the observed inverse relationship between coupling and translational strength, since coupling is expected to saturate once the mRNA is permanently unfolded. Whatever mechanism may be predominant, translational coupling was suggested to stabilize the stoichiometry of protein complexes and thus allows a better control of relative expression levels (109).

4.3.2. Selection for robustness can explain order of chemotaxis genes

Our experimental and computational analyzes show that in addition to the pathway topology and the co-transcription of genes, translational coupling is another mechanism contributing to the robustness of the chemotaxis system. The functional importance of such additional pairwise coupling of protein levels was shown by the enhancement of chemotaxis upon expression of any tested endogenous gene pair from a bicistronic construct compared to the overexpression of individual genes. In addition, correlation of protein levels on a single-cell level were found to be strongest for most efficiently spreading cells in a chemotaxis assay, demonstrating selection for enhanced posttranscriptional coupling. Thus translational coupling seems to improve the robustness of the output level CheYp and consequently of the CW motor bias against stochastic variations in translational levels.

Our in silico analysis showed that particular arrangement of chemotaxis genes can increases the robustness against translational noise. Specifically, those permutations that maximized number of gene pairs having opposing effects on the CheYp level were highest ranked. Although better knowledge of model parameters would be needed to resolve the exact positions of highest ranked permutations, gene order found in *E. coli* performed best under the assumption that the weakly translated genes *cheR* and *cheB* exhibit slightly increased noise levels compared to the stronger translated genes *cheY* and *cheZ*. Experiments and theoretical analysis thus suggest that gene order in the chemotaxis operons is not random, but was selected for maximizing coupling between expression of antagonistic proteins and thus increasing robustness of the pathway output.

Selection for coupling in all investigated pairs can be explained by known properties of the chemotaxis pathway. The kinase CheA and adaptor protein CheW are found in complexes with chemotaxis receptors (43, 125), whose stoichiometry and functional properties are influenced by relative levels of the corresponding proteins (79, 142). Hence it is not surprising that relative translation of CheA and CheW is strictly controlled. From the point of robustness, the coupling of CheY and CheZ can be well understood. An augmented level of CheY is counterbalanced by the increased concentration of the phosphatase CheZ thus ensuring homeostasis of the level of phosphorylated CheY. Coupling of CheB and CheR is expected to be beneficial, since these enzymes are antagonistically controlling the steady state level of receptor methylation and in turn the kinase activity of CheA. Thus ensuring a proper ratio of these two proteins contributes to the robustness of the CheYp output. Finally, the coupling of CheB and CheY is not surprising either. Both compete for stimulation dependent binding to the P2 domain of CheA and subsequent phosphorylation by this kinase (68, 82). In addition, higher CheB activity decreases the steady state level of receptor methylation and thus CheY phosphorylation. Hence a coelevated level of CheY, which would lead to an increase of CheYp on its own, is counteracted by reduction of CheA kinase activity due to increased demethylation of receptors. On the other side, upregulation of CheY should counterbalance an increase of CheB levels.

4.3.3. Evolution of gene order in chemotaxis operons

Consistent with our model predictions, a bioinformatic analysis revealed that there is a selection for preferred gene pairs rather than for the complete gene order (88). This makes it quite implausible that the observed consensus is the result of conservation or lateral transfer of the whole chemotaxis operon. Thus it is more likely that individual genes have been rearranged multiple times through evolution. Differences in gene orders of closely related species might thus reflect variations in network topology and gene regulation.

Our proposed mechanism of robustness- and noise-driven gene ordering in the chemotaxis operon can be regarded as a refinement of models explaining operon formation through selection for the coregulation of genes (120). Such a common control of genes may be beneficial if they encode components of the same pathway or a multiprotein complex. A closely related model is the balance hypothesis (111, 158). It postulates that an imbalance in the levels of two subunits of a multicomplex leads to the formation of complexes with wrong stoichiometry rendering them nonfunctional and are thus assumed to be under purifying selection. This hypothesis can be readily applied to metabolic operons, since they often encode components of larger enzyme complexes and such a selective pressure might have shaped the coupling of CheA and CheW. In contrast our model explaining the order of the other chemotaxis genes does not rely on the assumption of direct interaction or stable complex formation. Moreover, it predicts that coupling of certain genes is harmful and those might have undergone negative selection. We thus could expand the existing regulation based model of operon formation by elucidating possible causes explaining the internal structure of the chemotaxis operon.

However, our model does not account for the process of operon formation itself. Since bacterial genes which are more proximal to each other exhibit a stronger correlation of their expression (20, 64), the most likely mechanism is the gradual increase in proximity of favorable gene pairs. This would in turn lead to an increase of translational coupling and thus in robustness of the pathway output.

4.3.4. Conclusion

Our results reveal the importance of translational coupling and gene order for the robustness of chemotaxis signaling in *E. coli*. The preferential pairing of certain chemotaxis genes predicted by our computational model was confirmed by a bioinformatic analysis of different sequenced bacterial genomes. Such an organization of genes might be also evolutionary beneficial for other signaling networks by improving the robustness of the pathway output. Since it can be achieved without additional costs of the increased complexity it is expected to be common in bacterial networks.

5. Conclusion and outlook

Protein biosynthesis lies at the very heart of the life process and is of utmost importance to ensure the reproduction and sustainment of cells (113, 87). It is organized into two distinct phases, transcription and translation. Transcription yields the messenger RNA, an exact copy of the nucleotide sequence defining the gene to be expressed. In contrast, translation implies the interpretation of the sequence of nucleotides as a sequence of codons, i.e. triplets of nucleotides, thereby specifying the amino acids which have to be lined up for protein-assembly. This thesis investigated two different aspects of translation and the consequences for genome organization in *E. coli* and other bacteria. First we explained the deviation of codon usage at the beginning of genes found in many bacterial genomes. Second, we aimed to understand the selective advantage of the gene order in the chemotaxis *meche* operon. In the following we will give a short summary of the thesis and its main results. Subsequently we will discuss the relevance of our findings in a broader context and suggest possible future experiments.

Chapter 2 introduced the relevant background information about gene expression in bacteria with an emphasis on translation. There we also developed a simplified model of the translational process by breaking it down into different phases, i.e. initiation and elongation, but without resolving the movement of the ribosomes along the mRNA. Hence the model relies on ordinary differential equations and consequently a coarse grained description of translation. The model was subsequently used to analyze the effect of slowly translated codons at the gene start in chapter 3 and served as a starting point for the modeling of translational coupling in chapter 4.

In chapter 3 we investigated the possible causes for the observed deviation of codon usage at the gene start found across many bacterial genomes (42, 154). The current explanation for this phenomenon assumes a selection for slowly translated codons at the gene start in order to slow down early elongation, preventing "traffic jams" of ribosomes further downstream (154). According to the authors such a selection gives rise to a "ramp" of slowly translated codons at the beginning of genes. We found that the unusual codon usage strongly correlates with the suppression of mRNA structure around the translation start which is important in determining the translation efficiency of a gene (30, 31, 76, 124). This suggest that evolution selected such codons at the gene start which restrain the formation of mRNA structure around the ribosome binding site (RBS) and consequently allow an efficient translation initiation. In line with this "structure hypothesis" we found a strong decrease of the GC3-content and

even a moderate reduction of the GC1- and GC2-content content at the beginning of *E. coli* genes. To further test our hypothesis against the "ramp hypothesis" we differentiated within each genome between rare and abundant codons, from which the former should on average correspond to low abundant and the latter to high abundant tRNAs. Consistent with our hypothesis we found an increase of rare AU3 codons and a depletion of abundant GC3 codons at the gene start in *E. coli*. In contrast there was no change in the frequency of rare GC3 codons and a slight increase of abundant AU3 codons. When comparing many bacterial genomes, we found an enrichment of rare codons if they strongly overlapped with the set of AU3 codons and a corresponding depletion of abundant codons if they mainly consist of GC3 codons. Moreover, enrichment of rare and depletion of abundant codons depends on the global GC content of the genome, which readily can be explained by the "structure hypothesis": A high GC content of the transcripts implies an increased propensity to form stable secondary RNA structure, hence the pressure to suppress mRNA folding is stronger for such genomes. This in turn results in the choice of rare over abundant codons, as the former consist mainly of AU3 codons, which form on average less stable structures. In contrast, the "ramp hypothesis" cannot explain such a dependence on the GC content as it predicts the deviation of the codon usage at the beginning of genes to be an universal feature. Our hypothesis also can account for the observed asymmetry of the deviation of the GC3 content at gene start: There is almost no genome which shows an increase of the GC3 content, but especially genomes with a high global GC content exhibit a significant depletion of GC3 codons which is in line with a selection for suppression of mRNA structure. The GC content of a genome determines the composition of the set of rare and abundant codons, rendering rare codons GC poor in a genome with high GC content. Therefore the "ramp hypothesis" would predict a symmetric deviation of the local GC3 content with an increase for genomes with a low global GC content. Based on our hypothesis, we furthermore tested whether evolutionary simulations could reproduce the observed trends. Our algorithm optimized sequences of shuffled synonymous codons to resemble the average native folding by randomly exchanging synonymous codons at the gene start. Although the observed deviation of the codon usage and GC3 content was smaller than for the native sequences, our simulations could reproduce the observed correlation of unusual codon usage and suppression of structure around the translation start. Finally, we investigated experimentally the effects of codon usage and RNA structure at the gene start on translation efficiency. To this end we changed the folding energy while keeping the same codon usage at the beginning of two native *E. coli* genes. In addition, codon usage was altered while maintaining the same folding energy. In order to measure translation efficiency of these constructs by flow cytometry and qRT-PCR, we fused them to the $5'$ terminus of the yellow fluorescent protein gene. In agreement with our hypothesis, changing the structural level at the gene start markedly affects translation efficiency. In contrast, modifying the codon usage led to less conclusive result. To summarize, we supplied several arguments that selection for suppression of structure formation is

probably the most important driving force for differential codon usage at the gene start.

Another mechanism which may rely on folding of mRNA is translational coupling between adjacent open reading frames (ORFs) in an operon. The second part of the thesis (chapter 4) investigated the role of such coupling in the robustness of the chemotaxis pathway of *E. coli* and the consequences for the organization of genes in the *meche* operon. Our coworkers showed experimentally that translational coupling between chemotaxis genes indeed exists. To this end bicistronic constructs harboring pairs of chemotaxis genes as found in the genome were employed. The second gene was fused to an enhanced yellow fluorescent report gene (*eyfp*) to measure expression levels. By selectively adapting the RBS of the first open reading frame, whose efficiency was tested with a plasmid only carrying this gene also fused to *eyfp*, translational coupling could be demonstrated: An elevated expression of the second gene was observed for all constructs with an enhanced RBS of the upstream gene. In order to develop a framework to take into account translational coupling when simulating gene expression noise, we used the model for the translational process introduced in chapter 2 as a starting point. In addition, a model of the chemotaxis signaling system was employed to calculate the stationary, i.e. adapted pathway output. We made the assumption that translational coupling predominantly works downstream. Hence an increase in ribosome occupancy of an upstream ORF will raise the translation efficiency of the downstream gene. Translational coupling thus will lead to correlated fluctuations in protein levels between adjacent genes. Translational coupling may be a combination of several factors: Local depletion or increase of ribosome concentration, re-initiation of elongating ribosomes, or modulation of RNA structure around the RBS by ribosomes translating the upstream gene. However, the latter mechanism can account for the observation of coupling saturation which is expected once the structure is completely unfolded. From a functional point of view, such a mechanism may allow for a coupling between ORFs exhibiting different translational efficiencies, as modulating the accessibility of the RBS is equivalent to a change in the ribosome concentration. Indeed, expression levels of chemotaxis proteins differ by more than one order of magnitude (83). Certain pairs of these proteins have antagonistic effects on the adapted pathway output. For such pairs translational coupling thus implies that the overexpression of one gene will be partially compensated by the overexpression of the downstream gene. In contrast, pairs of genes encoding proteins having similar effects on the pathway output are avoided. This finding is in line with the experimental result that the overexpression of native pairs of chemotaxis genes is better tolerated than overexpression of single genes. To summarize, we provided experimental and theoretical evidence that translational coupling is important in minimizing the negative impact of gene expression noise by correlating translation of adjacent genes. Under the assumptions of our translational model, the most likely explanation for the observed gene order in the chemotaxis operon is thus that it maximizes the number of such pairs thereby being an important determinant for operon organization.

Both these projects highlight the importance of RNA structure in the process of transla-

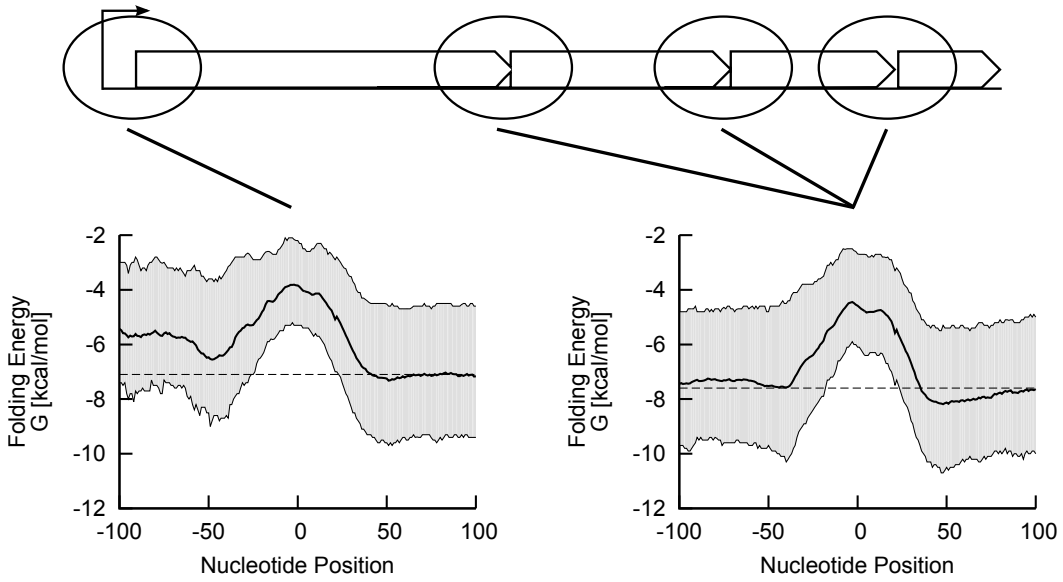


Figure 5.1.: mRNA is less structured in the 5′ UTR. Both figures show the mean energy and the inter-quartile range as grayed area for *E. coli* sequences aligned at the gene start. In the left figure only genes at the beginning of a transcriptional unit were used, whereas in the right figure gene starts within operons were aligned. The mRNA is less structured in the 5′ UTR than in the upstream region of genes within operons suggesting different selective pressures.

tion. The most likely explanation for the observed suppression of RNA structure around the translation start is that efficient translation initiation requires an unfolded mRNA. However, if translational coupling is mediated via the modulation of RNA structure by elongating ribosomes on the upstream gene, making the RBS of the downstream ORF more accessible, selection should favor certain structural elements in the RBS of such a gene. Hence the selection for efficient translation initiation independent of the upstream ORF and the selection for translational coupling should be opposed to each other. On an operon there is one site where selective forces should act in only one direction: The translation start site of genes at the beginning of transcriptional units. Indeed, as figure 5.1 shows, the 5′ UTR of mRNAs are less structured than the upstream regions of genes within operons. We therefore speculate that this effect might be due to an even more pronounced selection for efficient translation at the very beginning of transcriptional units than further downstream. However, to understand the relation between structural features of the mRNA and their effect on the translational process additional experiments would be necessary. One such experiment could be to express two fluorescent proteins from a bicistronic construct. This would make it possible to directly characterize cell-to-cell variations and to apply appropriate measures for correlations. We then could systematically change the translation efficiency of the first gene and modify regulating sequences between the two open reading frames as well as the distance between both genes. The prediction from our model would be that coupling should

be stronger the more compact folding is around the RBS of the downstream gene. This may impair initiation independent of the upstream gene and thus nature most probably had to find a compromise between translation efficiency and coupling.

To get a better understanding of the evolution of codon usage and RNA folding at the gene start we could perform our *in silico* evolutionary experiments *in vivo* by selecting for enhanced expression of genes artificially impaired in their translation efficiency. In such an experiment we could employ a system like the *lac* operon of *E. coli*, whose transcription can be externally controlled (115, 47). Then only the beginning of the *lac* genes has to be changed such that RNA structure around the RBS gets more stable thereby likely reducing translation efficiency of these genes. The selective pressure for an increased gene expression then could be simply switched on by replacing glucose by lactose in the growth medium. Moreover, the fixation of beneficial mutations could be directly monitored by measuring the growth rate. To characterize the evolutionary path best it would be preferable to sequence the whole genome during the course of the experiment. This way one can also detect whether increased gene expression really can be attributed to a change in translation efficiency or whether changes in the promoter or whole gene duplications have taken place. As an alternative to such an experiment, it would be also feasible to use cell sorting in order to directly select for enhanced expression levels of a gene fused to a fluorescent reporter.

Our understanding of the relation between folding and translation can also be expected to profit from the recent development of high-throughput methods. Thanks to next-generation sequencing techniques it has become possible to determine the transcriptome of a cell population quantitatively (100, 161, 94). Moreover, the ribosome footprint protocol allows to map the position of ribosomes along transcripts, thus being very useful to identify translational pausing sites (62, 61, 81). Combined with a novel method (parallel analysis of RNA structure - PARS) which determines folding profiles of whole transcriptomes it will become feasible to investigate the interdependence of RNA structure formation and translation on a genome-scale, not being restricted to certain study cases any more (70, 160).

From a practical point of view, it might be beneficial to take effects of RNA folding on translation into account when expressing heterologous genes. Usually an adaptation of the codon usage across the whole gene is considered to be of great advantage in such a case. As we have seen here, it is at least equally important to take into account the folding of mRNA around the RBS. There are additional relevant features in determining the translation efficiency like the structure of the Shine-Dalgarno sequence and efforts were undertaken to design artificial RBS (124). However, usually only upstream sequences are modified. Here we showed that downstream sequences do also play a role and thus codons should be synonymously substituted to suppress RNA structure in order to obtain the maximal yield of proteins. In addition, translational coupling could play a role for the expression of heterologous genes coding for protein complexes, but in order to exploit this in a predictive way we still have to gain a better understanding of both the folding characteristics of mRNA and its relation to translation.

A. Supplement for chapter 2

A.1. Effects of cell division on concentration and particle numbers

Dynamics of average concentration

If we are interested in the dynamics of concentration changes, we have to take into account cell divisions. We denote the cellular concentration of interest by c and calculate the total time derivative

$$\frac{dc}{dt} = \frac{\partial c}{\partial V}\frac{dV}{dt} + \frac{\partial c}{\partial t}, \tag{A.1}$$

where V denotes the considered volume. Since the concentration is by definition

$$c = \frac{n}{V}, \tag{A.2}$$

where n equals the particle number within the volume V, we can evaluate the first partial derivative on the right side of eq. (A.1) which gives us

$$\frac{\partial c}{\partial V} = -\frac{n}{V^2} = -\frac{c}{V}. \tag{A.3}$$

In order to proceed, we have to make an assumption about $V = V(t)$. For bacteria in the exponential phase the total volume of the population grows exponentially. If we just want to describe the average cellular concentrations we can think of the whole population as one volume, i.e. refer concentrations to the population volume. Since concentration is an intensive quantity it does not change when scaling down to the cell volume, however this is only valid if we neglect fluctuations of cellular concentrations. By doing so, we get for $V = V(t)$

$$V(t) = V_0 \exp(\gamma t), \tag{A.4}$$

where γ is the growth rate constant. Hence

$$\frac{dV}{dt} = \gamma V_0 \exp(\gamma t) = \gamma V(t). \tag{A.5}$$

Plugging this and eq. (A.3) into eq. (A.1) we obtain

$$\frac{dc}{dt} = -\gamma c + \frac{\partial c}{\partial t}. \tag{A.6}$$

In addition to volume growth concentrations also change due to chemical reactions and thus

$$\frac{\partial c}{\partial t} = \sum_{\rho} \nu_\rho w_\rho, \tag{A.7}$$

where ν_ρ are the number of moles produced $\nu_\rho > 0$ or consumed $\nu_\rho < 0$ in the reaction ρ which takes place with the velocity w_ρ (105). Equation (A.7) might also take into account active degradation. We therefore finally obtain our result

$$\frac{dc}{dt} = \sum_{\rho} \nu_\rho w_\rho - \gamma c. \tag{A.8}$$

Dynamics of copy numbers

We now want to adopt the perspective of the single cell again. Let us assume that we have n copies of a given molecule, for example a protein. Every cell division will halve the number of copies, if we neglect fluctuations due to an asymmetric distribution of molecules to the daughter cells. The effect of cell division on copy numbers can be best understood if we assume that translation is halted. Then the copy number after l cell divisions reads

$$n(l) = n_0 \left(\frac{1}{2}\right)^l. \tag{A.9}$$

This can be restated with time t as independent variable by using $l = t/\tau$, where τ is the generation time

$$n(t) = n_0 \left(\frac{1}{2}\right)^{t/T}. \tag{A.10}$$

For times $t \gg \tau$ this process is approximately continuous and we can take the time derivative yielding

$$\frac{dn}{dt} = -\frac{\ln(2)}{\tau} n(t). \tag{A.11}$$

Since the growth rate constant γ and the generation time τ are related by

$$\gamma = \frac{\ln(2)}{\tau}, \tag{A.12}$$

we obtained the same result as with eq. (A.6).

A.2. Algorithmic prediction of RNA secondary structure

We denote the sequence S of N nucleotides by

$$S = s_1, s_2, \ldots, s_N, \tag{A.13}$$

indexed in the $5' \rightarrow 3'$ direction of the RNA. The s_i denote the single nucleotides, i.e.

$$s_i \in \{\text{A,U,C,G}\}. \tag{A.14}$$

The RNA secondary structure contains stacked base-pairs and loops. Only the canonical base-pairs A:U, G:C and the wobble base-pair G:U are allowed. The RNA secondary structure of R can then be described as an undirected graph $G = (V, E)$ with the nucleotides as vertices V and the edges E being either nucleotides connected by ester bounds s_{i-1}, s_i, s_{i+1} or base-pairs $s_i : s_j = i : j, 1 \leq i < j <\leq N$ (144). Two base-pairs $i : j$ and $k : l$ have to obey one of the restrictions

$$i = k \text{ and } j = l \tag{A.15}$$

$$i < j < k < l \tag{A.16}$$

$$i < k < l < j. \tag{A.17}$$

The first restriction implies the identity of the two base-pairs. The other two restrictions can be best understood by using a rainbow diagram or circle plot. Here, two nucleotides forming a base-pair are connected with an arc. Thus the second restriction can be depicted in such a diagram as two arcs next to each other. This means that there can be more than one hairpin, each base-pair belonging to another stem. The third forbids crossing arcs, thus excluding pseudo-knots. A fourth restriction is imposed by the minimal size of a loop, i.e. for a base-pair $i : j$

$$j - i \geq 4 \tag{A.18}$$

has to hold (144). These restrictions allow to apply a recursive scheme to determine the structure with minimal free energy. For illustration's sake we will show how such a dynamic programming algorithm works, however solving the more simple problem of maximizing the number of base-pairs instead of minimizing free energy, which was first solved by Nussinov and coworkers (107, 37). Consider the fragment $S(i, j)$ between nucleotide i and j of sequence S. We now want to find the maximal number of base-pairs $M(i, j)$ which can be formed with this sequence. To keep things simple, we relax restriction (A.18) and allow loops of zero length. First let us evaluate the case, that the nucleotide i base-pairs with j. Due to restriction (A.17) $M(i + 1, j - 1)$ is independent of the formation of a base-pair $i : j$, hence

$$M(i, j) = 1 + M(i + 1, j - 1). \tag{A.19}$$

If i or j do not form a base-pair, we get

$$M(i, j) = 0 + M(i + 1, j) \tag{A.20}$$

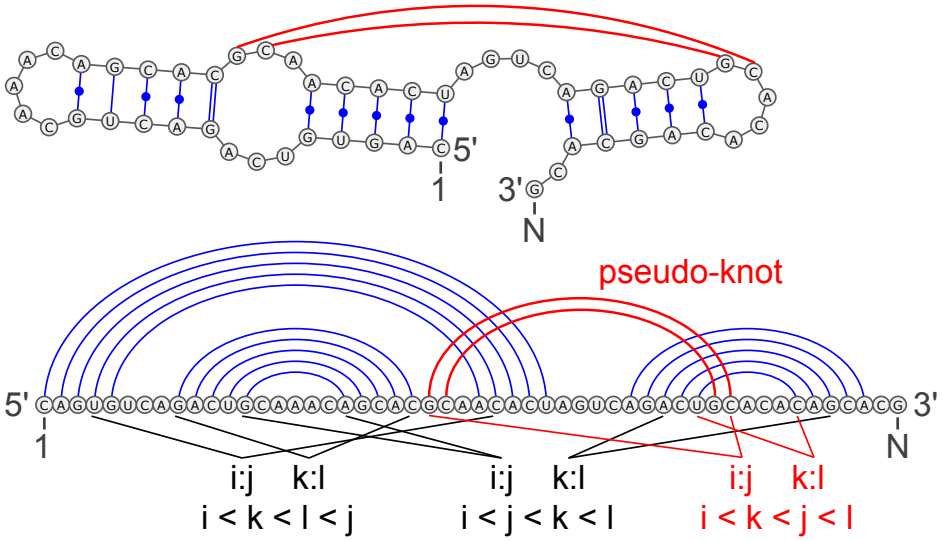


Figure A.1.: At top of the figure the common representation of a RNA secondary structure plus pseudo-knot is shown, whereas below the rainbow diagram is used to depict the same structure. The base-pairs represented by the blue archs respect the restrictions (A.15) - (A.18), but the red archs, stemming from a pseudo-knot, cross the blue ones and therefore lead to a violation of these restrictions.

in the first, and

$$M(i,j) = 0 + M(i,j-1) \tag{A.21}$$

in the second case. Finally i and j may base-pair but with other nucleotides of the sequence $S(i+1, j-1)$. Due to (A.16) and (A.17) this will give us two independent substructures, thus

$$M(i,j) = \max_{i<k<j} M(i,k) + M(k+1,j). \tag{A.22}$$

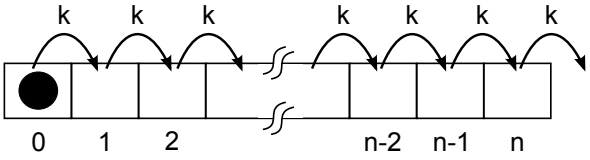
Altogether we therefore obtained a recurrence relation

$$M(i,j) = \max \begin{cases} 1 + M(i+1,j-1) & \text{if } i,j \text{ pair} \\ M(i+1,j) & \\ M(i,j-1) & \\ \max_{i<k<j} M(i,k) + M(k+1,j). & \end{cases} \tag{A.23}$$

Hence the problem has been reduced to a smaller problem. The algorithm fills the upper $N \times N$ triangular matrix defined by $M(i,j)$. The matrix is initialized by $M(i,i) = M(i,i-1) = 0$ and then the diagonals are filled by iteration using eq. (A.23). The algorithm stops once $M(1,N)$, i.e. the maximal number of base-pairs which can be formed within the sequence S, is determined. However, we still have to construct the structure corresponding to $M(1,N)$. This is accomplished by tracing back the path which brought us by maximizing the number of base-pairs into the corner $(1,N)$ (37).

A.3. Analytic solution of a simple elongation model

Assume a particle hopping in discrete steps with an uniform constant k in one direction along an array of length $L = n + 1$, finally hopping off when moving out of position n. We



can interpret this as a toy model of the elongating ribosome in the limit of low ribosome occupancy of the mRNA and equal elongation rate constants k. We now want to determine the probability p_i to find the ribosome at position i by solving the $n+1$ ordinary differential equations (ODEs)

$$\frac{d}{dt}p_0 = -kp_0 \tag{A.24}$$

$$\frac{d}{dt}p_i = k(p_{i-1} - p_i), \quad 1 \leq i \leq n, \tag{A.25}$$

together with the initial conditions

$$p_0(0) = 1 \tag{A.26}$$

$$p_i(0) = 0, \quad 1 \leq i \leq n. \tag{A.27}$$

This means that at the beginning the particle sits at position 0 with certainty. Using the Laplace transform turns this system of ODEs into a system of linear equations,

$$s\tilde{p}_0 - 1 = -k\tilde{p}_0 \tag{A.28}$$

$$s\tilde{p}_i = k(\tilde{p}_{i-1} - \tilde{p}_i), \quad 1 \leq i \leq n, \tag{A.29}$$

where

$$\tilde{p}_i(s) = \int_0^\infty e^{-st} p_i(t) dt \tag{A.30}$$

is the Laplace transform of $p_i(t)$, defined for complex s (16). This set of equations can be easily solved, yielding

$$\tilde{p}_i = \frac{k^i}{(s+k)^{i+1}}, \quad 0 \leq i \leq n. \tag{A.31}$$

The inverse transformation can be best accomplished by applying the residue theorem (17) which gives us

$$p_i(t) = \frac{e^{-kt}(kt)^i}{i!}, \quad 0 \leq i \leq n. \tag{A.32}$$

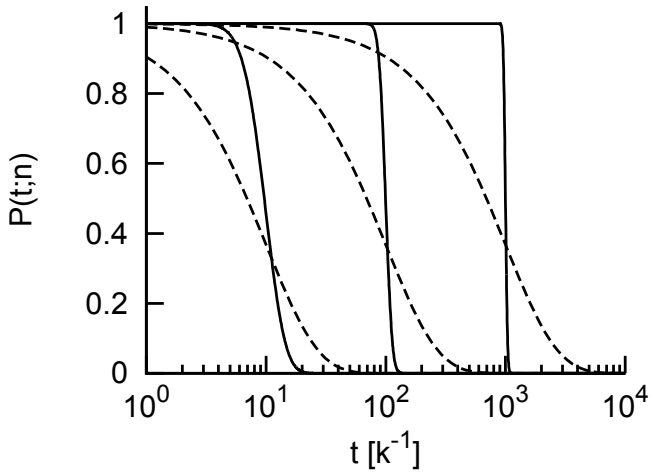


Figure A.2.: The figure shows the exact solution $P(t;n) = e^{-kt}\sum_{i=0}^{n}\frac{(kt)^i}{i!}$ as solid line and the approximation $P^*(t;n) = e^{-k_e t}$ as dashed line. The effective rate constant $k_e = \frac{k}{n+1}$ was determined according to eq. (A.39). Shown are plots for $L = n+1 = 10, 100, 1000$ from the left to the right. In contrast to the exact solution, the approximation shows no delay but summing up the $n+1$ rate constants in eq. (A.39) interpolates the exact behavior.

The probability $P(t;n)$ that the particle is still bound to the array is given by

$$P(t;n) = e^{-kt}\sum_{i=0}^{n}\frac{(kt)^i}{i!}. \tag{A.33}$$

If the array is infinitely long, i.e. $n \to \infty$, $P(t;n)$ equals 1 for all times since

$$e^{kt} = \sum_{i=0}^{\infty}\frac{(kt)^i}{i!}. \tag{A.34}$$

The probability $P(t;n)$ starts to decline when $p_n(t)$ has become maximal at time $t_{\max}$. This time can be determined by solving

$$\frac{d}{dt}p_n(t) = 0 \tag{A.35}$$

for t which yields

$$t_{\max} = \frac{n}{k}. \tag{A.36}$$

To leave the array, the particle has to move one more step which takes about $\tau \sim k^{-1}$. Hence the overall characteristic time Δt the particle is bound to the array is given by

$$\Delta t = \frac{n+1}{k}. \tag{A.37}$$

Since the particle has to move in total $n+1$ steps, each with a characteristic timescale $\tau \sim k^{-1}$, we recover the result given by eq. (2.12).

We now want to approximate this process by a single step process with an effective rate

$$\frac{d}{dt}P^*(t;n) = -k_e P^*(t;n), \tag{A.38}$$

with k_e given by

$$k_e = \frac{1}{\Delta t} = \frac{k}{n+1}. \tag{A.39}$$

We used the star $*$ to indicate the approximation. The initial condition reads

$$P^*(0;n) = 1. \tag{A.40}$$

Hence we get

$$P^*(t;n) = e^{-\frac{kt}{n+1}}. \tag{A.41}$$

We compared the approximation $P^*(t;n)$ as given by eq. (A.41) and the exact solution $P(t;n)$, given by eq. (A.33) in figure A.2 for different n.

B. Supplement for chapter 3

B.1. Details of bioinformatics analysis

Sequence database

Genomic data for *E. coli* was obtained from the set of flat files of the EcoCyc database (71) in version 13.6. The data for the investigated 414 bacterial genomes was extracted from the flat files provided by the BioCyc database (66) in version 13.6. Tier3 database was used, generated by the PathoLogic program, to predict operons amongst others. We skipped the first gene of each transcriptional unit (TU) in our analysis since the respective $5'$ untranslated regions (UTR) may not be well annotated in most genomes. Non-chromosomal genes were not analyzed, and only protein coding sequences were included in the analysis. For genes present in multiple TUs, only the largest TU for those genes was taken into account. Furthermore, we discarded TUs containing genes of nucleotide length which are not multiple of 3 or which contained components that were incompletely annotated or could not be found in the database. Finally, splicable genes (that are genes with programmed frameshifts) were omitted from the analysis. Taxonomy was derived using the NCBI taxonomy ids provided through the BioCyc database and the Perl module Bio-LITE-Taxonomy-NCBI-0.08.

Null models

We used two null models: One with shuffled codons (SC), where codons are randomly permuted within a single gene, and another which preserves the amino acid sequence by only shuffling synonymous codons (SSC). Codons with sequencing errors were not shuffled. The start and stop codons and also sequences of overlapping genes were always preserved.

Calculation of folding energy

We used the Vienna RNA Package, version 1.8.5 available at `http://www.tbi.univie.ac.at/%7Eivo/RNA/`, to predict free energy of RNA sequences (59). Gibbs free energy was calculated within a sliding window of size $2 \times 19 + 1 = 39$ nucleotides which is suggested as the approximate number of nucleotides covered by a ribosome (13). Each calculated value was assigned to the nucleotide position of the window center.

Codon frequencies and Kullback-Leibler divergence

For each set of synonymous codons, we determined the global frequency $q_{i,j}$ of each codon within this set. The index i indicates the amino acid and runs from 1 to 20, and j indexes the synonymous codon (running from 1 to the number of synonymous codons S_i, which ranges from 1 to maximum 6). Additionally position-dependent codon frequencies were determined, $p_{i,j}(k)$, with k denoting codon position relative to the translation start site, where the start codon (typically AUG) corresponds to $k = 0$. In both cases, pseudocount regularization is applied by adding a pseudocount of 1. Using these frequencies, we calculated the position dependent Kullback-Leibler divergence KLD(k) that quantifies the deviation of the codon usage downstream of the translation start codon

$$\mathrm{KLD}(k) = \sum_{i=1}^{20} \sum_{j=1}^{S_i} p_{i,j}(k) \ln \frac{p_{i,j}(k)}{q_{i,j}}.$$

We used the global frequencies $q_{i,j}$, determined from native sequences, for all calculations. Theoretically, this measure is zero if both distributions are identical. Due to finite size effects, the Kullback-Leibler divergence KLD(k) is biased to values larger than zero (55, 122). The used estimator for the used Kullback-Leibler divergence is given by

$$\hat{\mathrm{KLD}}(k) = \sum_{i=1}^{20} \sum_{j=1}^{S_i} \hat{p}_{i,j}(k) \ln \frac{\hat{p}_{i,j}(k)}{\hat{q}_{i,j}}, \tag{B.1}$$

where S_i denotes the number of synonymous codons and

$$\hat{p}_{i,j}(k) = \frac{n_{i,j}(k)}{N_i(k)} \tag{B.2}$$

$$\hat{q}_{i,j} = \frac{m_{i,j}}{M_i}. \tag{B.3}$$

are the estimators for the frequencies at position k relative to the gene start and for the global frequencies of codon j coding amino acid i without pseudocount regularization, respectively. The corresponding observed counts are denoted by $n_{i,j}(k)$ and $m_{i,j}$. The total counts are given by

$$N_i(k) = \sum_{j=1}^{S_i} n_{i,j}(k) \tag{B.4}$$

$$M_i = \sum_{j=1}^{S_i} m_{i,j}. \tag{B.5}$$

Since $M_i \gg N_i(k)$ the finite size effect stems predominantly from the estimate of $p_{i,j}(k)$ by $\hat{p}_{i,j}(k)$ and therefore we assume $\hat{q}_{i,j} \approx q_{i,j}$. We introduce a smallness parameter $\varepsilon_{i,j}(k)$ to

quantify the deviation of the estimator $\hat{p}_{i,j}(k)$ from the frequency $p_{i,j}(k)$

$$\hat{p}_{i,j}(k) = p_{i,j}(k)(1 + \varepsilon_{i,j}(k)), \tag{B.6}$$

i.e.

$$\varepsilon_{i,j}(k) = \frac{\hat{p}_{i,j}(k) - p_{i,j}(k)}{p_{i,j}(k)}. \tag{B.7}$$

We now expand eq. (B.1) in powers of $\varepsilon_{i,j}(k)$ up to quadratic order and obtain

$$\begin{aligned} \hat{\mathrm{KLD}}(k) &\approx \sum_{i=1}^{20}\sum_{j=1}^{S_i}\left[p_{i,j}(k)\ln\frac{p_{i,j}(k)}{q_{i,j}}\right. \\ &+ \left. p_{i,j}(k)\left(1+\ln\frac{p_{i,j}(k)}{q_{i,j}}\right)\varepsilon_{i,j}(k) + p_{i,j}(k)\frac{\varepsilon_{i,j}^2(k)}{2} + \mathcal{O}(\varepsilon_{i,j}^3(k))\right]. \end{aligned} \tag{B.8}$$

In order to obtain the bias of the Kullback-Leibler divergence we calculate the mean of eq. (B.8). To this end, we need to know $\langle\varepsilon_{i,j}(k)\rangle$ and $\langle\varepsilon_{i,j}^2(k)\rangle$. The codon (i,j) at position k occurs by definition with probability $p_{i,j}(k)$ and the probability that it does not occur is given through normalization by $1 - p_{i,j}(k)$. The number of occurrence therefore follows a binomial distribution and we obtain

$$\langle n_{i,j}(k)\rangle = N_i(k)p_{i,j}(k) \tag{B.9}$$

$$\langle\langle n_{i,j}^2(k)\rangle\rangle = \langle n_{i,j}^2(k)\rangle - \langle n_{i,j}(k)\rangle^2 = N_i(k)p_{i,j}(k)(1-p_{i,j}(k)) \tag{B.10}$$

and thus

$$\langle\varepsilon_{i,j}(k)\rangle = 0 \tag{B.11}$$

$$\langle\varepsilon_{i,j}^2(k)\rangle = \frac{1-p_{i,j}(k)}{N_i(k)p_{i,j}(k)}. \tag{B.12}$$

Using these results, we obtain

$$\langle\hat{\mathrm{KLD}}(k)\rangle \approx \mathrm{KLD}(k) + \sum_{i=1}^{20}\sum_{j=1}^{S_i} p_{i,j}(k)\frac{1-p_{i,j}(k)}{2N_i(k)p_{i,j}(k)} + \mathcal{O}\left(\frac{1}{N_i^2(k)}\right) \tag{B.13}$$

$$= \mathrm{KLD}(k) + \sum_{i=1}^{20}\sum_{j=1}^{S_i}\frac{1-p_{i,j}(k)}{2N_i(k)} + \mathcal{O}\left(\frac{1}{N_i^2(k)}\right) \tag{B.14}$$

$$= \mathrm{KLD}(k) + \sum_{i=1}^{20}\frac{S_i-1}{2N_i(k)} + \mathcal{O}\left(\frac{1}{N_i^2(k)}\right), \tag{B.15}$$

where we used the normalization of $p_{i,j}(k)$, i.e.

$$\sum_{j=1}^{S_i} p_{i,j}(k) = 1. \tag{B.16}$$

Enrichment of extreme codons

Abundant and rare codons are by our definition the 15 most abundant and 15 most rare codons, measured by their codon frequencies. The total frequencies of rare f_{frare} and abundant f_{abund} codons are the sum over the codon frequencies in the corresponding set, i.e.

$$f_{\text{rare}} = \sum_{i \in \{\text{rare codons}\}} f_i$$

where f_i denotes the frequency of the ith codon and the same definition applies for abundant codons. In addition, we calculated this frequency at each position downstream of the start codon and defined the fold change fc of total codon frequency at position k as

$$\text{fc}_{\text{rare}}(k) = \frac{f_{\text{rare}}(k)}{f_{\text{rare}}},$$

and correspondingly for abundant codons. In order to quantify an enrichment of these extreme codons at the beginning of genes, we calculated the average fold change $\overline{\text{fc}}$ from position 1 through 5 downstream of the translation start codon. Using the null model SSC, we calculated empirical Z values, using $n = 100$ realizations of the null model,

$$Z_{\overline{\text{fc}}} = \frac{\overline{\text{fc}} - \langle \overline{\text{fc}}_{\text{nm}} \rangle}{\sqrt{\langle \langle \overline{\text{fc}}^2_{\text{nm}} \rangle \rangle}},$$

where $\langle \overline{\text{fc}}_{\text{nm}} \rangle = \frac{1}{n} \sum_{i=1}^{n} \overline{\text{fc}}_{\text{nm}i}$ and $\langle \langle \overline{\text{fc}}^2_{\text{nm}} \rangle \rangle = \frac{1}{n-1} \sum_{i=1}^{n} \left(\overline{\text{fc}}_{\text{nm}i} - \langle \overline{\text{fc}}_{\text{nm}} \rangle \right)^2$.

Simulated evolution

We evolved whole genomes consisting of M genes from the *E. coli* genome, whose synonymous codons, except those coding for the translation start, were initially shuffled (SSC null model). Sequencing errors were removed by replacing them with randomly chosen bases out of the subsets of bases defined by the IUPAC code. Additionally 100 of the 414 bacterial genomes were evolved in the same way. Furthermore, only non overlapping genes were taken into account, leading to a total number of $M = 1088$ genes and in the case of the 100 other genomes ranging from 114 to 1471 genes (on an average of 690 ± 323 genes). Each of the first 26 codons, excluding the start codon, was mutated with a probability $p_m = 10^{-5}$ by substituting it with a synonymous codon, which was chosen with a frequency according to the global codon usage.

Fitness of each genome was determined using the deviation of the folding energy of each sequence from the native mean folding energy. The nucleotides which were considered in the calculation of the folding energy range from $N_u = 32$ to $N_d = 51$ nucleotides relative to the translation start site. To accelerate the simulation we did not calculate the folding energy at each position but used a step size of $\Delta n = 9$, leading to a total number of $N = 6$ energy calculations. Subsequent averaging yielded the used fitness measure f

$$f = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=0}^{N-1} \frac{[G(p(n)) - G_m(p(n))]^2}{\sigma(p(n))^2},$$

where $\sigma(p(n))$ and $G(p(n))$ denote the standard deviation and average of folding energy of native sequences at the position $p(n) = n\Delta n + w_h - N_u$, respectively, and $G_m(p(n))$ indicates the folding energy of the evolved sequence m at the corresponding position.

The genetic algorithm used tournament selection and elitism to determine which genomes are selected. Populations of 1000 genomes were evolved in each case, from which an elite of the 100 best were selected in each round that were not mutated subsequently. In contrast genomes chosen by tournament selection were mutated.

Evolution is terminated, once the relative change of the maximal fitness in the population within 100 generation is lower than 0.002% ($|f_{t-100} - f_t|/|f_{t-100} + f_t| < 10^{-5}$).

In the case of *E. coli* the simulation was performed 100 times from the same shuffled genome, whereas for the 100 other bacteria the simulation was started only once per genome. In the latter case, the native baseline energy was used to calculate ΔG of the null model and evolved sequences.

B.2. Experimental details

Evaluation of synonymous sequences

For each cytoplasmic protein, we generated *in silico* all synonymous sequences that differ in the first 6 codons. Thereby only the beginning of the genes, i.e. $5'$ UTR and 31 codons including ATG, was taken into account. To select appropriate genes for disentangling codon usage and folding energy we had to judge each synonymous sequences according to this measures. We calculated the mean folding energies around the translation start (i.e. folding energies of 39 nucleotide (nt) long stretches centered at nucleotide positions $0, \pm 3, \pm 6$ relative to the translation start). Codon adaptation was estimated by determining the geometric mean of the relative tRNA abundance corresponding to codons from position 1 to 6 (34, 169, 170). The folding energy and codon usage profiles for the selected genes *ypdE* and *pykA* are shown in fig. B.4.

Plasmids and strains

We designed for each gene five constructs (*wt*, *minCA*, *maxCA*, *minG*, *maxG*) with an 50 nt long native 5′ UTR in which additional ATG codons were removed to avoid alternative start sites, the start codon ATG, the stretch of variable codons, a sequence of 24 unaltered codons, followed by three codons coding for glycine (see fig. 3.12b in chapter 3). These constructs were synthesized de novo (Eurofins MWG Operon) and delivered in high copy plasmids (pCR2.1, pEX-A). We cloned them directly or after PCR amplification via the restriction sites XbaI and NcoI into an *yfp* containing pETDuet-1 vector thereby yielding a fusion protein with YFP. Restriction enzymes, Pfu polymerase and corresponding buffers were purchased from Fermentas. Primers were ordered from metabion. For plasmid isolation GeneJET Plasmid Miniprep Kit and for DNA cleanup GeneJET PCR Purification Kit was used (both from Fermentas). For gel elution NucleoSpin Extract II Kit (Macherey-Nagel) was used.

The target vector is a gift of the working group of Ilka Axmann (Charité Berlin). They cloned *yfp* (sequence in the supplement) via the restriction sites NcoI and BamHI into an empty pETDuet-1 vector (Novagen). All obtained plasmids carry therefore a T7 promoter-1 inducible by β-D-thiogalactoside (IPTG) and an ampicilin resistance. Plasmids were transformed into an *E. coli* BL21(D3) strain.

Growth conditions

Overnight cultures were grown in Lysogeny broth (LB; 10 g/l tryptone 5 g/l yeast extract, 10 g/l NaCl) containing ampicilin (100 μg/ml) at 37°C. For measurements of YFP expressions, overnight cultures were diluted 1:50 in 20 ml fresh LB containing ampicilin. They were grown in a rotary shaker to an optical density at 600 nm (OD_{600}) of approximately 0.5($\pm$0.1) and then induced with IPTG at a concentration of 45 μM. One hour after induction cell cultures were diluted 1:5 in 10 ml fresh LB containing ampicilin and IPTG at the indicated concentration, grown for one more hour and then harvested by centrifugation (7 min, 900 g, 4°C), decanting of the supernatant and the resuspended in tethering buffer (5 mM K_2HPO_4, 5 mM KH_2PO_4 0.1 mM EDTA, 1 μM L-methionine, 0.1 % (v/v) lactic acid [pH 7]). Resuspended cells were stored at approximately 4°C for around one hour before measurements. Samples qRT-PCR were also harvested by centrifugation (7 min, 2500 g, RT). After decanting the supernatant samples for qRT-PCR were frozen in liquid nitrogen and those for qRT-PCR samples were stored at -80°C.

Quantification of gene expression

Flow cytometry measurements Median expression levels of fluorescent proteins were quantified in a population of approximately 10^5 cells using flow cytometry on a FACSCalibur (BD Biosciences) equipped with an argon 488-nm laser. Measurement was triggered by forward

ypdE	nucleotide sequence
5′ UTR and start codon	GAAGTGCTCTACGCCAAGCCGAAAACAGTGTTGCTCACGG GAGAGGCATAATG
wt	GAT TTA TCG CTA TTA AAA
minG	GAC CTC TCC CTC CTA AAA
maxG	GAT CTA TCT TTA CTT AAA
minCA	GAT CTA AGT CTA CTA AAG
maxCA	GAT CTG AGC TTG CTG AAA
3′ sequence and glycine linker	GCGTTGAGCGAGGCAGATGCGATCGCCTCCTCGGAACAGG AAGTGCGGCAGATCCTGCTGGAAGAAGCGGATGGCGGCGGA
pykA	
5′ UTR and start codon	GGATTTCAAGTTCAAGCAACACCTGGTTGTTTCAGTCAAC GGAGTATTACATG
wt	TCC AGA AGG CTT CGC AGA
minG	AGT AGG CGG CTC CGT AGG
maxG	TCA AGA AGA TTA CGC AGA
minCA	TCA CGA CGA CTA CGG AGG
maxCA	TCT CGT CGT CTG CGT CGT
3′ sequence and glycine linker	ACAAAAATCGTTACCACGTTAGGCCCAGCAACAGATCGCG ATAATAATCTTGAAAAAGTTATCGCGGCGGGTGGCGGCGGA

Table B.1.: Sequences derived from *ypdE* and *pykA*. *wt*, *minG*, *maxG*, *minCA* and *maxCA* indicate the corresponding altered sequences.

scatter (fsc) and sideward scatter (ssc) events. FACSCalibur data files were imported into MATLAB by using "fca_readfcs.m" (developed by Laszlo Balkay from the University of Debrecen) and then analyzed in MATLAB. Only non-zero measurements were taken into account for fsc and ssc values between the 10th and 90th percentile. Median value of the autofluorescence background, measured for control cells, was subtracted from all values.

qRT-PCR measurements Total RNA was isolated from pelleted cells using the InviTrap Spin Cell RNA Mini Kit (Stratec Molecular) following the protocol for Gram-negative bacteria. The RNA quality was judged by the absorbance ratio A_{260nm}/A_{280nm}. Samples of isolated RNA were treated with DNAseI (Fermentas) for one hour at 37°C to remove remaining traces of DNA. DNAseI was inactivated by incubating the samples with 50 mM EDTA at 65°C for 10 minutes.

cDNA was produced from total RNA using Reverse Transcription Kit (Fermentas) with

random hexamer primer. For each sample a control was performed without Revert Aid Transcriptase (RT-). mRNA was quantified by qRT-PCR with gene specific primers on a real-time PCR cycler (7500 Fast Real-Time PCR System, Applied Biosystems) using Fast SYBR Green Master Mix (Applied Biosystems). The level of *yfp* mRNA was normalized to the level of *gapdh* mRNA as an internal standard. In all cases we got significantly higher ct values for the RT- control (for *yfp* primers $>$ 10 ct, for *gapdh* primers $>$ 7 ct). Two clear outliers were removed from the analysis.

B.3. Supplementary figures

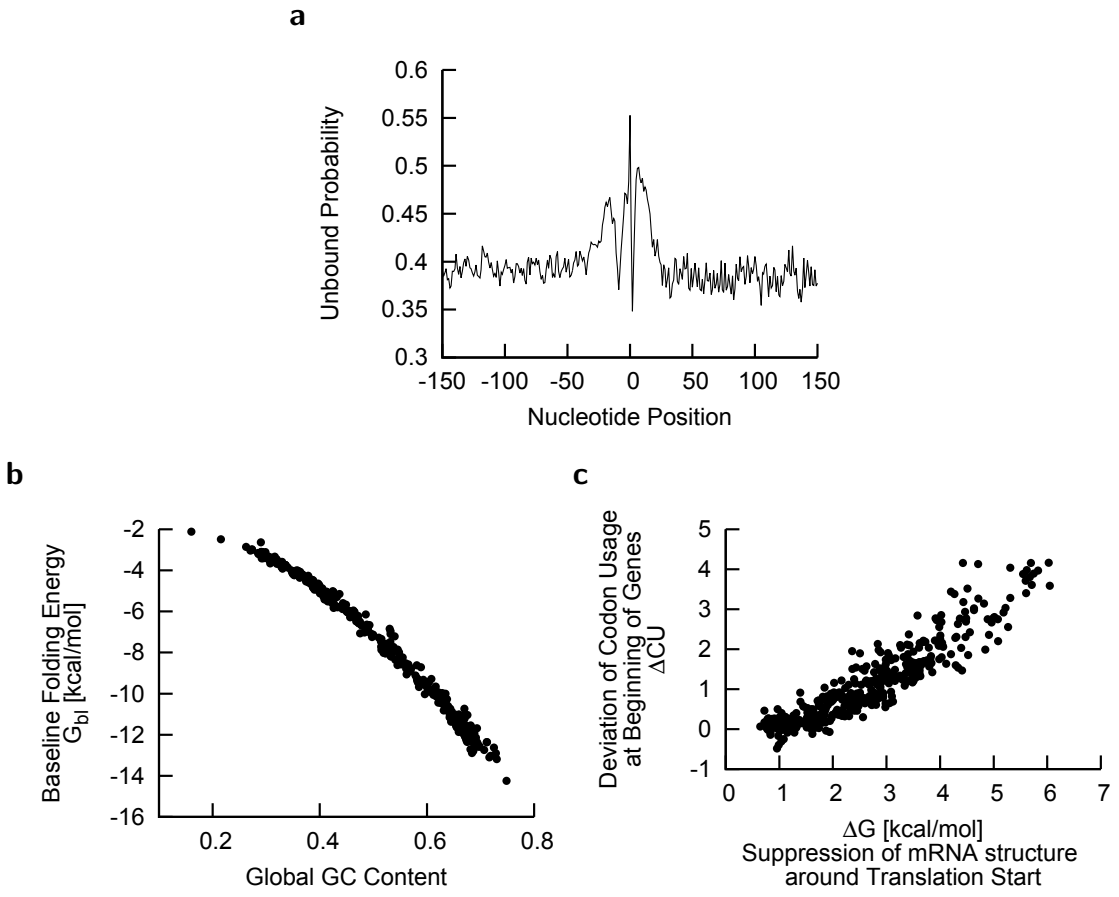


Figure B.1.: (**a**) The average probability to remain unpaired increases around the gene start. This probability was calculated for each sequence spanning a range from -160 to +160 nucleotides relative to the gene start and averaged subsequently. (**b**) The baseline folding energy G_{bl} depends on the global GC-content of a genome: The higher the GC-content the more stable mRNA folds on average. (**c**) Correlation of suppression of mRNA structure around translation start and deviation of codon usage remains when the analysis is restricted to non-overlapping genes.

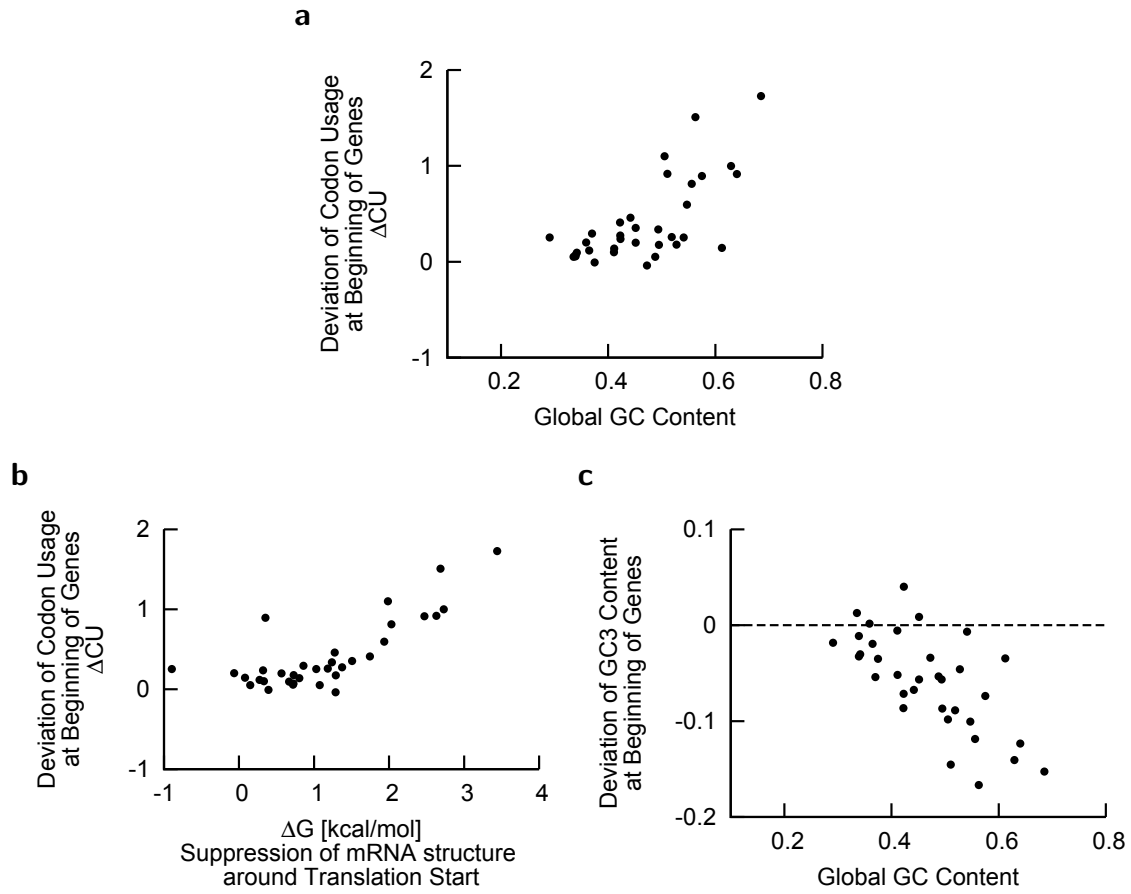


Figure B.2.: Archaea genomes exhibit similar features as bacterial genomes. All major findings for bacterial genomes also hold for archaea. A total of 34 genomes were investigated. (**a**) The deviation of codon usage at gene start, ΔCU is stronger the higher the global GC-content and consequently the more stable mRNA folds on average. This result is the same as for bacterial genomes, shown in fig. 3.5a. (**b**) As for bacterial genomes (see figs. 3.5b and B.1c), the deviation from usual codon usage, ΔCU correlates with suppression of mRNA secondary structure around gene start, ΔG (correlation coefficient $r = 0.78$). (**c**) Archaea genomes also show an asymmetry in deviation from global GC3-content near translation start site. GC3-content is almost only decreased and the stronger the higher the genomic GC-content. The observed correlation coefficient $r = -0.68$ is similar to the one obtained for bacterial genomes (see fig. 3.10b).

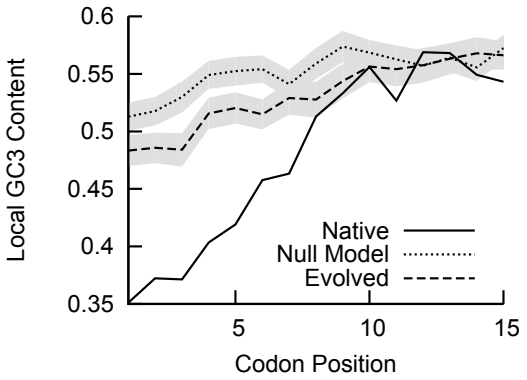


Figure B.3.: GC3-content at the gene start in evolved genome. Deviation from genomic GC3-content near the start codon of evolved sequences differs from the null model (SSC) and shows similar functional behavior as sequences taken from the *E. coli* genome.

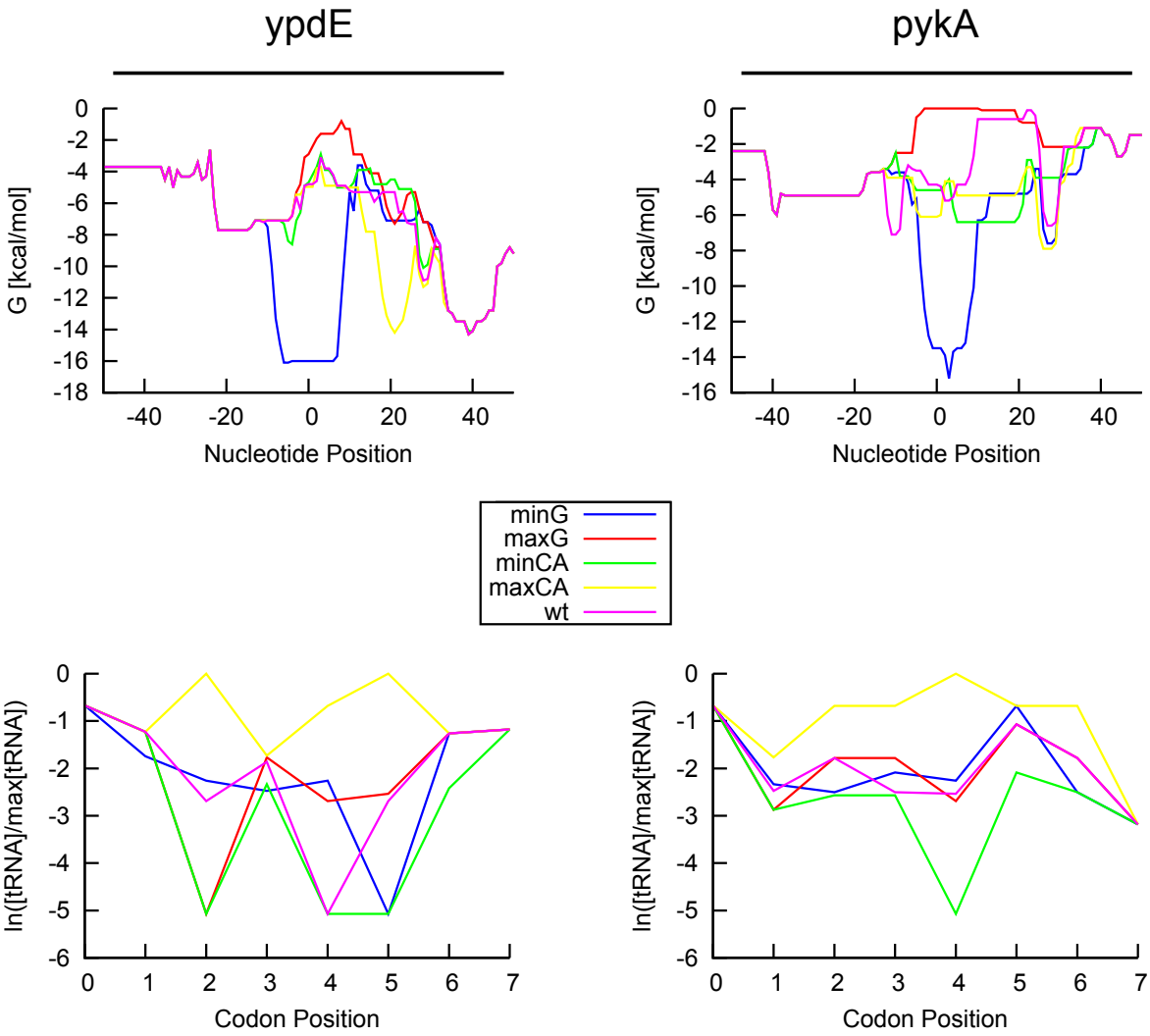


Figure B.4.: Folding energy and codon usage profiles for investigated constructs. The constructs *minG* and *maxG* exhibit different degree of structure at the gene start, but have on average the same codon usage as the *wt* sequence. In contrast, mRNA folding energies of *minCA* and *maxCA* do not differ much from the *wt* construct, but show opposed codon adaptation.

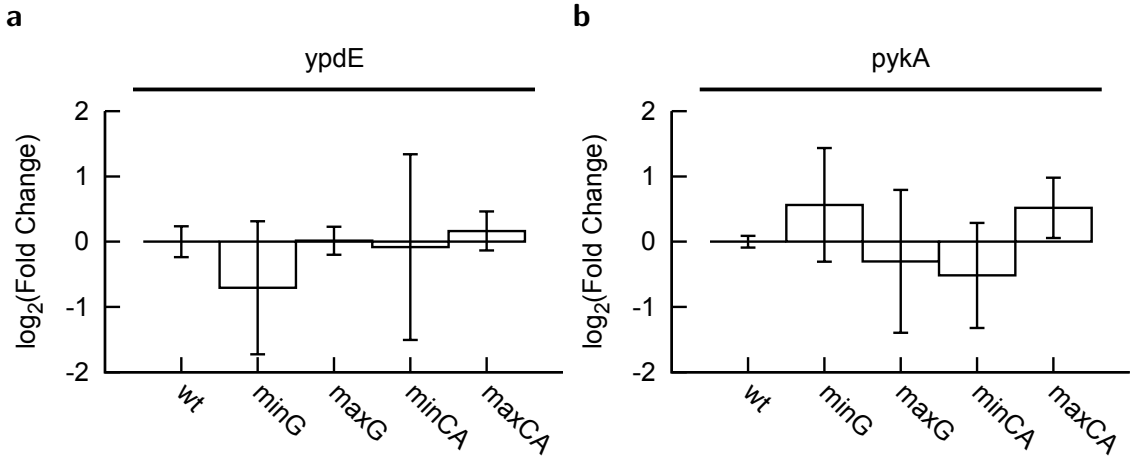


Figure B.5.: Results of qRT-PCR measurements for the different constructs. Within the experimental error mRNA levels as determined by qRT-PCR do not differ between constructs derived from the respective *E. coli* genes.

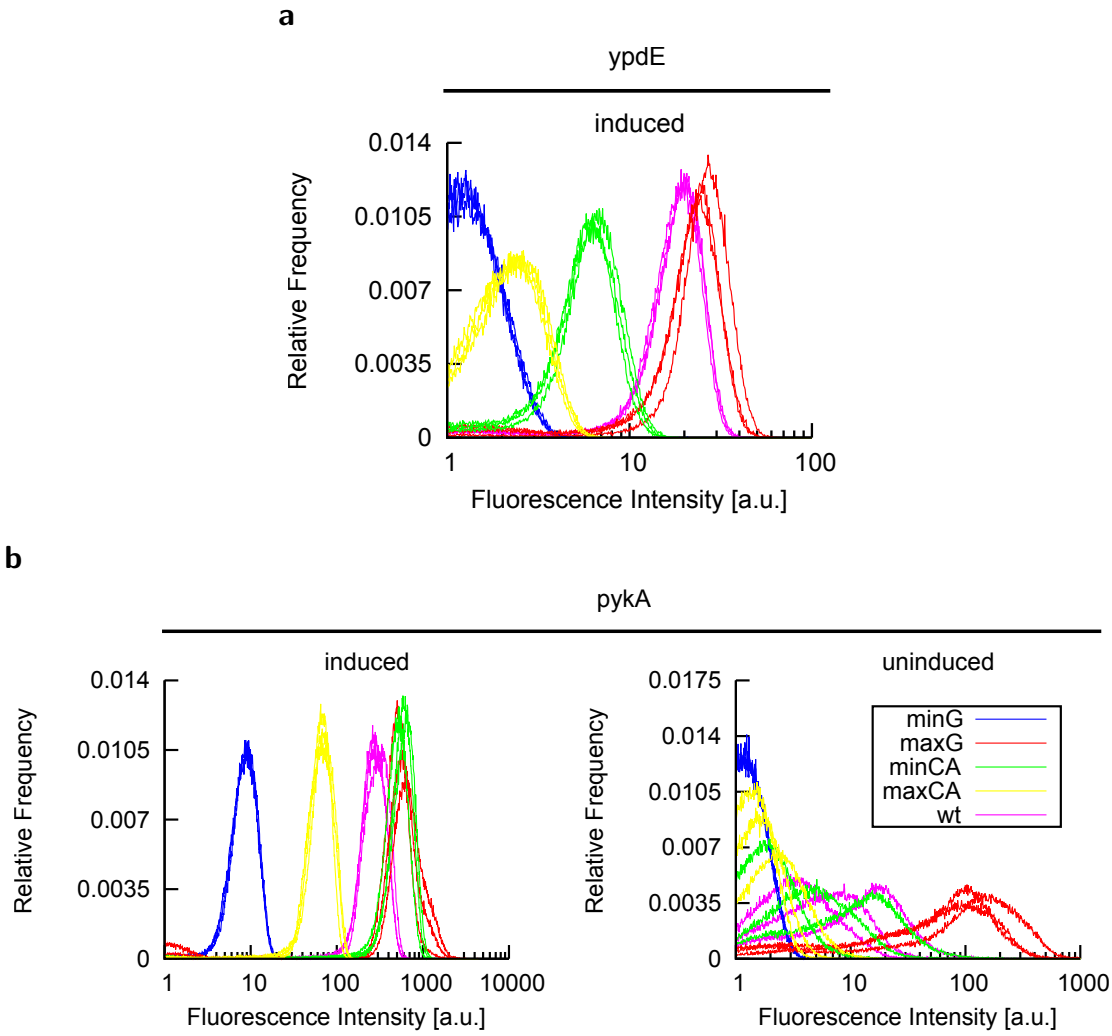


Figure B.6.: Distribution of fluorescence levels of investigated constructs as measured by flow cytometry. Values are background corrected.

C. Supplement for chapter 4

C.1. Experimental details

Strains and plasmids

Escherichia coli K-12 strains used in this study were derived from RP437 (112). All strains and plasmids are summarized in table C.1 and C.2. Monocistronic constructs expressing YFP fusions to CheR, CheB, CheY, CheZ and CheA under moderately strong ribosome binding sites (RBSs) and pTrc promoter inducible by isopropyl β-D-thiogalactoside (IPTG) have been described before (89, 68, 84, 139, 141). They were used to obtain constructs with strong RBSs (summarized in table C.3) and bicistronic constructs by using PCR and cloning to modify the upstream sequence. Because expression of *cheY* is strongly upregulated by a sequence inside *cheB* gene (A.M. and V.S., unpublished), a non-translated 316 nucleotide fragment of *cheB* was included upstream of the *cheY* start codon in pVS319 (-316_*cheY-eyfp*) plasmid to achieve expression comparable to pVS142 (*cheB_cheY-eyfp*) construct. To reduce levels of expression for the *cheB_cheY-eyfp* and -316_*cheY-eyfp* constructs, both fragments were cloned under weaker pBAD promoter inducible by L-arabinose, to obtain pLL33 and pLL36, respectively.

Strain	Description/Relevant Genotype	Reference
RP437	wild type for chemotaxis	(112)
VS100	Δ*cheY*	(139)
VS104	Δ*cheYcheZ*	(141)
VS161	Δ*cheZ*	(89)
RP4972	Δ*cheB*	J.S. Parkinson, personal gift

Table C.1.: Strains used to study the role of translational coupling in robustness of bacterial chemotaxis pathway.

Plasmid	Description	Reference
pTrc99A	Expression vector; pBR ori, pTrc promoter, Amp^R	(4)

Continued on next page

Plasmid	Description	Reference
pBAD33	Expression vector; pACYC ori, pBAD promoter, Cm^R	(52)
pDK57	RBS^{CheYS2}_$CheA_S$-YFP expression plasmid; pTrc99A derivate	(68)
pDK66	Expression vector for cloning of C-terminal YFP fusions; RBS^{CheYS} pTrc99A derivative	(69)
pVS18	RBS^{CheY}_CheY-YFP expression plasmid; pTrc99A derivate	(141)
pVS64	RBS^{CheZ}_CheZ-YFP expression plasmid; pTrc99A derivate	(84)
pVS88	RBS^{CheY}_CheY-YFP_ RBS^{CheZ}_CheZ-YFP bicistronic construct; pTrc99A derivate	(142)
pVS137	RBS^{CheR}_CheR-YFP expression plasmid; pTrc99A derivate	(89)
pVS138	RBS^{CheB}_CheB-YFP expression plasmid; pTrc99A derivate	(89)
pVS142	RBS^{CheB}_CheB_CheY-YFP expression plasmid; pTrc99A derivate	this work
pVS145	RBS^{CheR}_CheR_CheB-YFP expression plasmid; pTrc99A derivate	this work
pVS261	RBS^{CheYS}_CheA-YFP expression plasmid; pTrc99A derivate	this work
pVS305	RBS^{CheY}_CheY_CheZ-YFP expression plasmid; pTrc99A derivate	this work
pVS319	-316_CheY-YFP expression plasmid; pTrc99A derivate	this work
pVS321	$RBS^{CheY\uparrow}$_CheY_CheZ-YFP expression plasmid; pTrc99A derivate	this work
pVS450	$RBS^{CheB\uparrow}$_CheB_CheY-YFP expression plasmid; pTrc99A derivate	this work
pVS451	$RBS^{CheR\uparrow\uparrow}$_CheR_CheB-YFP expression plasmid; pTrc99A derivate	this work
pVS452	$RBS^{CheR\uparrow\uparrow}$_CheR-YFP expression plasmid; pTrc99A derivate	this work
pVS487	$RBS^{CheB\uparrow}$_CheB-YFP expression plasmid; pTrc99A derivate	this work
pVS490	RBS^{CheYS2}_CheA_CheW-YFP expression plasmid; pTrc99A derivate	this work

Continued on next page

Plasmid	Description	Reference
pVS495	$RBS^{CheY\uparrow}$_CheY-YFP expression plasmid; pTrc99A derivate	this work
pVS520	RBS^{CheYS}_$CheA_S$_CheW-YFP expression plasmid; pTrc99A derivate	this work
pAM80	$RBS^{CheR\uparrow}$_CheR-YFP expression plasmid; pTrc99A derivate	this work
pAM81	$RBS^{CheR\uparrow}$_CheR_CheB-YFP expression plasmid; pTrc99A derivate	this work
pLL33	-316_CheY-YFP expression plasmid; pBAD33 derivate	this work
pLL36	RBS^{CheB}_CheB_CheY-YFP expression plasmid; pBAD33 derivate	this work

Table C.2.: Plasmids used to study the role of translational coupling in robustness of bacterial chemotaxis pathway. See table C.3 for description and exact sequence of RBS.

Construct	Upstream Sequence
RBS^{CheR}	*GAGCTC*TTGAGAAGGCGCT**ATG**
RBS^{CheB}	*GAGCTC*AGTAAGGATTAACG**ATG**
RBS^{CheY}	*GAGCTC*CGTATTTAAATCAGGAGTGTGAA**ATG**
RBS^{CheZ}	*GAGCTC*CAGGGCATGTGAGGATGCGACT**ATG**
RBS^{CheYS}	*ACTAGT*GAAGGAGTGTGCC**ATG**
$RBS^{CheR\uparrow}$	*GAGCTC*GATAGGGTGGGCGCT**ATG**
$RBS^{CheR\uparrow\uparrow}$	*GAGCTC*GATAGGAAAGGCGCT**ATG**
$RBS^{CheB\uparrow}$	*GAGCTC*AAGAGGAAATTAACG**ATG**
$RBS^{CheY\uparrow}$	*GAGCTC*AATAGAGGAAATGTGAA**ATG**

Table C.3.: Upstream ribosome binding sequences of the fusion constructs. Italic type indicates recognition site of restriction enzymes, SacI or SpeI, used for cloning constructs, boldface font indicates the start codon. A single upward arrow (↑) indicates an enhanced RBS, and double arrows (↑↑) indicate a strongly enhanced RBS.

Growth conditions

Overnight cultures were grown in tryptone broth (TB; 1% tryptone, 0.5% NaCl) containing ampicillin (100 μg/ml) or chloramphenicol (100 μg/ml) at 30°C for 16 hours. For mea-

surements of the YFP expression in liquid cultures, overnight cultures were diluted 1:100 in fresh TB containing ampicillin and indicated concentrations of IPTG or L-arabinose. Cell cultures were allowed to grow 3.5 – 4 hours at 34°C in a rotary shaker until OD_{600} of 0.45, then harvested by centrifugation (8000 rpm, 1 min), washed and resuspended in tethering buffer (10 mM potassium phosphate, 0.1 mM EDTA, 1 μM L-methionine, 10 mM sodium lactate [pH 7]).

TB soft agar (swarm) plates were prepared by supplementing TB with 0.3% agar (Applichem), required antibiotics (100 μg/ml ampicillin; 34 μg/ml chloramphenicol), and indicated concentrations of IPTG and L-arabinose. Plates were inoculated using fresh cells from LB agar plates and swarm assays were performed for 6-8 hours at 34°C. Images of swarm plates were taken using a Canon EOS 300 D (DS6041) camera, and analyzed with ImageJ (Wayne Rasband, NIH) to determine the diameter of the swarm rings.

Quantification of gene expression

Mean expression levels of fluorescent proteins were quantified in a population of 10^4 cells as described before (73) using flow cytometry on a FACScan (BD Biosciences) equipped with an argon 488 nm laser. FACScan data were analyzed using CellQuestTM Pro 4.0.1 software. Mean value of the autofluorescence background, measured for control cells, was subtracted from all values. Single-cell protein levels were measured using fluorescence microscopy on a Zeiss AxioImager Z1 microscope equipped with ORCA AG CCD Camera (Hamamatsu) and HE YFP (Excitation BP 500/25; Dichroic LP 515; Emission BP 535/30) and HE CFP (Excitation BP 436/25; Dichroic LP 455; Emission BP 480/40) filter sets. Integral levels of fluorescence in individual cells were quantified using an automated custom-written ImageJ plug-in (89) and normalized to cell length to obtain relative concentrations of fluorescent proteins (152).

C.2. Modeling details

Receptor model

Consider homodimers of two-state receptors, being either active or inactive (41, 72). The receptor can bind a ligand in the active state

$$\mathrm{T^{on}} + \mathrm{L} \underset{k_-^{\mathrm{on}}}{\overset{k_+^{\mathrm{on}}}{\rightleftharpoons}} [\mathrm{T^{on}L}]$$

or the inactive state

$$\mathrm{T^{off}} + \mathrm{L} \underset{k_-^{\mathrm{off}}}{\overset{k_+^{\mathrm{off}}}{\rightleftharpoons}} [\mathrm{T^{off}L}] \,.$$

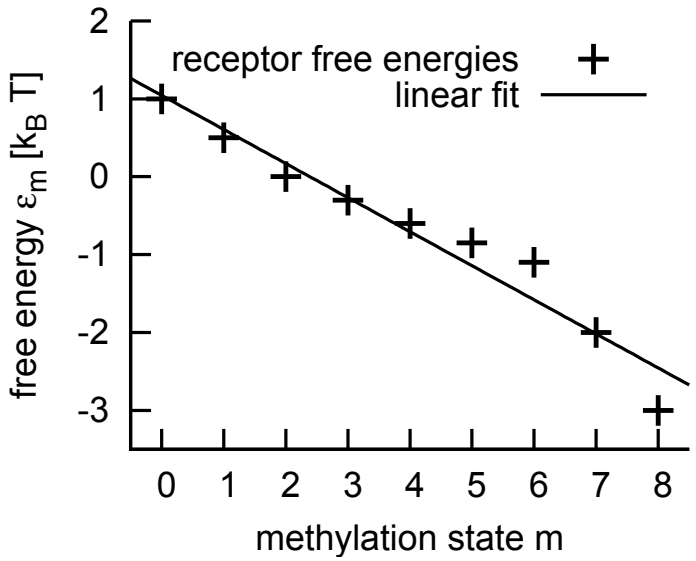


Figure C.1.: The free energy values ϵ_m as a function of the methylation state m of a receptor homo-dimer. The free energy is to good approximation a linear function of the methylation state. Data is taken from (41).

If we assume mass kinetics and detailed balance, we get

$$k_+^{\text{on}} T^{\text{on}} L - k_-^{\text{on}} [T^{\text{on}} L] = 0 \tag{C.1}$$
$$k_+^{\text{off}} T^{\text{off}} L - k_-^{\text{off}} [T^{\text{off}} L] = 0, \tag{C.2}$$

and this can be rewritten in terms of the probability $p(A, B)$, where $A = \{\text{on}, \text{off}\}$ denotes the activity-state and $B = \{\text{b}, \text{u}\}$ the binding state. Thus

$$p(\text{on}, \text{u}) L - K^{\text{on}} p(\text{on}, \text{b}) = 0 \tag{C.3}$$
$$p(\text{off}, \text{u}) L - K^{\text{off}} p(\text{off}, \text{b}) = 0, \tag{C.4}$$

where $K^i = k_-^i / k_+^i$ denotes the dissociation constant. In thermal equilibrium these probabilities are proportional to the Boltzmann factor, i.e.

$$p(A, B) \sim e^{-F(A,B)}, \tag{C.5}$$

where $F(A, B)$ is the corresponding free energy, measured in units of $k_B T$. Hence, by using (C.3) we get

$$F(\text{on}, \text{b}) = F(\text{on}, \text{u}) - \ln(L/K^{\text{on}}) = F_{\text{on}} - \ln(L/K^{\text{on}}), \tag{C.6}$$

and correspondingly

$$F(\text{off}, \text{b}) = F(\text{off}, \text{u}) - \ln(L/K^{\text{off}}) = F_{\text{off}} - \log(L/K^{\text{off}}). \tag{C.7}$$

The probability of the receptor to be in the on state, irrespective of the binding-state, therefore reads

$$p(\mathrm{on}) = \frac{e^{-[F_{\mathrm{on}} - \log(L/K^{\mathrm{on}})]} + e^{-F_{\mathrm{on}}}}{e^{-[F_{\mathrm{on}} - \log(L/K^{\mathrm{on}})]} + e^{-F_{\mathrm{on}}} + e^{-[F_{\mathrm{off}} - \log(L/K^{\mathrm{on}})]} + e^{-F_{\mathrm{off}}}}. \tag{C.8}$$

By introducing the offset free energy $\epsilon_m = F_{\mathrm{on}} - F_{\mathrm{off}}$, which depends only on the methylation state of the receptor, we can write (C.8) as

$$p(\mathrm{on}|m; L) = \frac{1}{1 + e^{\epsilon_m} \left(\frac{1+L/K^{\mathrm{off}}}{1+L/K^{\mathrm{on}}}\right)}, \tag{C.9}$$

and interpret as the conditional probability of the receptor to be in the on-state given a certain methylation level m. The ligand concentration L is a external parameter in this description. It is assumed, that binding favors the off state, i.e. $K^{\mathrm{off}} \ll K^{\mathrm{on}}$, whereas energy decreases with methylation, thus favoring the on state.

Phosphorylation of response regulator CheY

he full set of equations describing phosphorylation and dephosphorylation according to the reaction schemes (4.15) and (4.18) reads

$$\begin{aligned}\frac{d}{dt}Y &= -k_Y^1 Y(A[P_2|P_1] + A[P_2|P_1p]) + k_Y^2(A[P_2Y|P_1] + A[P_2Y|P_1p]) \\ &\quad - k_Y Y(A[P_2|P_1p] + A[P_2Yp|P_1p] + A[P_2Y|P_1p]) + k_Z^3[YpZ] \end{aligned} \tag{C.10}$$

$$\begin{aligned}\frac{d}{dt}Yp &= k_Y^5(A[P_2Yp|P_1p] + A[P_2Yp|P_1]) - k_Y^4 Yp(A[P_2|P_1] + A[P_2|P_1p]) \\ &\quad + k_Y Y(A[P_2|P_1p] + A[P_2Yp|P_1p] + A[P_2Y|P_1p]) - k_Z^1 YpZ + k_Z^2[YpZ] \end{aligned} \tag{C.11}$$

$$\frac{d}{dt}[YpZ] = k_Z^1 YpZ - (k_Z^2 + k_Z^3)[YpZ] \tag{C.12}$$

$$\frac{d}{dt}Z = -k_Z^1 YpZ + (k_Z^2 + k_Z^3)[YpZ] \tag{C.13}$$

$$\begin{aligned}\frac{d}{dt}A[P_2|P_1] &= -k_A p_A A[P_2|P_1] + k_Y Y A[P_2|P_1p] - k_Y^1 Y A[P_2|P_1] \\ &\quad + k_Y^2 A[P_2Y|P_1] + k_Y^5 A[P_2Yp|P_1] - k_Y^4 Yp A[P_2|P_1] \end{aligned} \tag{C.14}$$

$$\begin{aligned}\frac{d}{dt}A[P_2|P_1p] &= k_A p_A A[P_2|P_1] - k_Y Y A[P_2|P_1p] - k_Y^1 Y A[P_2|P_1p] \\ &\quad + k_Y^2 A[P_2Y|P_1p] + k_Y^5 A[P_2Yp|P_1p] - k_Y^4 Yp A[P_2|P_1p] \end{aligned} \tag{C.15}$$

$$\frac{d}{dt}A[P_2Y|P_1] = k_Y^1 Y A[P_2|P_1] - k_Y^2 A[P_2Y|P_1] - k_A p_A A[P_2Y|P_1] + k_Y Y A[P_2Y|P_1p] \tag{C.16}$$

$$\begin{aligned}\frac{d}{dt}A[P_2Y|P_1p] &= k_Y^1 Y A[P_2|P_1p] - k_Y^2 A[P_2Y|P_1p] + k_A p_A A[P_2Y|P_1] \\ &\quad - k_Y^3 A[P_2Y|P_1p] - k_Y Y A[P_2Y|P_1p] \end{aligned} \tag{C.17}$$

$$\begin{aligned}\frac{d}{dt}A[P_2Yp|P_1] &= k_Y^3 A[P_2Y|P_1p] - k_A p_A A[P_2Yp|P_1] + k_Y Y A[P_2Yp|P_1p] \\ &\quad - k_Y^5 A[P_2Yp|P_1] + k_Y^4 Yp A[P_2|P_1] \end{aligned} \tag{C.18}$$

$$\frac{d}{dt}A[P_2Yp|P_1p] = k_A p_A A[P_2Yp|P_1] - k_Y Y A[P_2Yp|P_1p] - k_Y^5 A[P_2Yp|P_1p] + k_Y^4 Yp A[P_2|P_1p] \tag{C.19}$$

Due to the small amount of CheB, i.e. $B^T \ll A^T$ and $B^T \ll Y^T$, all interactions of CheB with CheA are neglected in the set of differential equations governing the dynamics of CheA and CheY forms.

Phosphorylation of response regulator CheB

The differential equations for the phosphorylation of CheB are similar to those for CheY, however we assume that there is no re-binding to the P_2 domain after phosphorylation, i.e.

$k_B^4 = 0$. In contrast to CheY, the phosphorylated CheB auto-dephosphorylates by a rate γ_B.

$$\frac{d}{dt}B = -k_B^1 B(A[P_2|P_1] + A[P_2|P_1p]) + k_B^2(A[P_2B|P_1] + A[P_2B|P_1p]) - k_B B(A[P_2|P_1p] + A[P_2Y|P_1p] + A[P_2Yp|P_1p]) + \gamma_B Bp \tag{C.20}$$

$$\frac{d}{dt}Bp = k_B^5(A[P_2Bp|P_1p] + A[P_2Bp|P_1]) + k_B B(A[P_2|P_1p] + A[P_2Y|P_1p] + A[P_2Yp|P_1p]) - \gamma_B Bp \tag{C.21}$$

$$\frac{d}{dt}A[P_2B|P_1] = k_B^1 BA[P_2|P_1] - (k_B^2 + k_A p_A)A[P_2B|P_1] + k_Y Y A[P_2B|P_1p] \tag{C.22}$$

$$\frac{d}{dt}A[P_2B|P_1p] = k_B^1 BA[P_2|P_1p] + k_A p_A A[P_2B|P_1] - (k_B^2 + k_B^3 + k_Y Y)A[P_2B|P_1p] \tag{C.23}$$

$$\frac{d}{dt}A[P_2Bp|P_1] = k_B^3 A[P_2B|P_1p] - (k_A p_A + k_B^5)A[P_2Bp|P_1] + k_Y Y A[P_2Bp|P_1p] \tag{C.24}$$

$$\frac{d}{dt}A[P_2Bp|P_1p] = k_A p_A A[P_2Bp|P_1] - (k_Y Y + k_B^5)A[P_2Bp|P_1p]. \tag{C.25}$$

We neglect terms $k_B BA[P_2B|P_1p]$ and $k_B BA[P_2Bp|P_1p]$, since $k_B B \ll k_Y Y$ and $A[P_2B|P_1p] + A[P_2Bp|P_1p] \ll A[P_2|P_1p] + A[P_2Y|P_1p] + A[P_2Yp|P_1p]$.

Stationary state equations

We now want to derive equations which are valid in the stationary state. First, we introduce the state variables

$$A[P_2Y] = A[P_2Y|P_1] + A[P_2Y|P_1p] \tag{C.26}$$

$$A[P_2Yp] = A[P_2Yp|P_1] + A[P_2Yp|P_1p] \tag{C.27}$$

$$A[P_2] = A[P_2|P_1] + A[P_2|P_1p]. \tag{C.28}$$

Adding the corresponding eqs. (C.16) + (C.17) and (C.18) + (C.19), respectively, gives us

$$\frac{d}{dt}A[P_2Y] = k_Y^1 A[P_2]Y - k_Y^3 A[P_2Y|P_1p] - k_Y^2 A[P_2Y] \tag{C.29}$$

$$\frac{d}{dt}A[P_2Yp] = k_Y^4 A[P_2]Yp + k_Y^3 A[P_2Y|P_1p] - k_Y^5 A[P_2Yp]. \tag{C.30}$$

In the stationary state, we get

$$A[P_2Y] = \frac{k_Y^1}{k_Y^2}A[P_2]Y - \frac{k_Y^3}{k_Y^2}A[P_2Y|P_1p] \tag{C.31}$$

$$A[P_2Yp] = \frac{k_Y^4}{k_Y^5}A[P_2]Yp + \frac{k_Y^3}{k_Y^5}A[P_2Y|P_1p]. \tag{C.32}$$

The conservation of all $A[P_2]$ domains and the eqs. (C.31) and (C.32) give us

$$A[P_2] = A^T - A[P_2Y] - A[P_2Yp] \tag{C.33}$$

$$\approx \frac{A^T}{1 + Y/K_Y^D + Yp/K_{Yp}^D}, \tag{C.34}$$

where $K_Y^D = k_Y^2/k_Y^1, K_{Yp}^D = k_Y^5/k_Y^4$. The approximation becomes exact if $k_Y^2 = k_Y^5$ holds.

Using the definitions introduced above, we can rewrite eq. (C.17) for $A[P_2Y|P_1p]$,

$$\frac{d}{dt}A[P_2Y|P_1p] = k_Y^1 A[P_2|P_1p]Y + k_A p_A(A[P_2Y] - A[P_2Y|P_1p]) - \left(k_Y^2 + k_Y^3 + k_Y Y\right) A[P_2Y|P_1p]. \tag{C.35}$$

Solving for $A[P_2Y|P_1p]$ in the stationary state and using eq. (C.31), we get

$$A[P_2Y|P_1p] = \Omega_Y + \Gamma_Y A[P_2|P_1p] \tag{C.36}$$

$$\Omega_Y = \frac{\frac{k_Y^1}{k_Y^2} k_A p_A A[P_2]Y}{k_A p_A(1 + \frac{k_Y^3}{k_Y^2}) + k_Y^3 + k_Y Y + k_Y^2} \tag{C.37}$$

$$\Gamma_Y = \frac{k_Y^1 Y}{k_A p_A(1 + \frac{k_Y^3}{k_Y^2}) + k_Y^3 + k_Y Y + k_Y^2} \tag{C.38}$$

Similarly, we can formulate eq. (C.19) for $A[P_2Yp|P_1p]$ as

$$\frac{d}{dt}A[P_2Yp|P_1p] = k_Y^4 A[P_2|P_1p]Yp + k_A p_A(A[P_2Yp] - A[P_2Yp|P_1p]) - (k_Y^5 + k_Y Y)A[P_2Yp|P_1p]. \tag{C.39}$$

This gives us, using eqs. (C.32) and (C.36), $A[P_2Yp|P_1p]$ in the stationary state,

$$A[P_2Yp|P_1p] = \Omega_{Yp} + \Gamma_{Yp} A[P_2|P_1p] \tag{C.40}$$

$$\Omega_{Yp} = \frac{\frac{k_A p_A}{k_Y^5}\left(k_Y^3 \Omega_Y + k_Y^4 A[P_2]Yp\right)}{k_Y^5 + k_A p_A + k_Y Y} \tag{C.41}$$

$$\Gamma_{Yp} = \frac{\frac{k_A p_A}{k_Y^5} k_Y^3 \Gamma_Y + k_Y^4 Yp}{k_Y^5 + k_A p_A + k_Y Y}. \tag{C.42}$$

Finally, rewriting eq. (C.15) for $A[P_2|P_1p]$,

$$\frac{d}{dt}A[P_2|P_1p] = k_A p_A(A[P_2] - A[P_2|P_1p]) - A[P_2|P_1p]((k_Y^1 + k_Y)Y + k_Y^4 Yp) + k_Y^2 A[P_2Y|P_1p] + k_Y^5 A[P_2Yp|P_1p] \tag{C.43}$$

and solving for $A[P_2|P_1p]$ in the stationary state, we get

$$A[P_2|P_1p] = \frac{k_A A[P_2]p_A + k_Y^2 A[P_2Y|P_1p] + k_Y^5 A[P_2Yp|P_1p]}{k_A p_A + k_Y^1 Y + k_Y Y + k_Y^4 Yp}. \tag{C.44}$$

Using eqs. (C.36) and (C.40) yields

$$A[P_2|P_1p] = \frac{k_A p_A A[P_2] + k_Y^2 \Omega_Y + k_Y^5 \Omega_{Yp}}{k_A p_A + (k_Y^1 + k_Y)Y + k_Y^4 Yp - k_Y^2 \Gamma_Y - k_Y^5 \Gamma_{Yp}}. \tag{C.45}$$

Hence we can calculate $A[P_2|P_1p], A[P_2Y|P_1p]$ and $A[P_2Yp|P_1p]$ as a function of Y and Yp alone. The total amount of phosphorylated P_1 domains is simply given by

$$A[P_1p] = A[P_2|P_1p] + A[P_2Y|P_1p] + A[P_2Yp|P_1p]. \tag{C.46}$$

Free CheY can be calculated from a conservation law

$$Y = Y^T - Yp - [YpZ] - A[P_2Y] - A[P_2Yp]. \tag{C.47}$$

We make the approximation $k_Y^2 \approx k_Y^5$ and use eq. (C.34),

$$Y^T - Yp - [YpZ] - Y - \frac{A^T}{1 + Y/K_Y^D + Yp/K_{Yp}^D}\left(\frac{Y}{K_Y^D} + \frac{Yp}{K_{Yp}^D}\right) = 0. \tag{C.48}$$

This leads to an equation of second order in Y,

$$Y^2 + Y\overbrace{\left(A^T - Y^T + Yp + [YpZ] + K_Y^D\left(1 + \frac{Yp}{K_{Yp}^D}\right)\right)}^{b} + \underbrace{\frac{K_Y^D}{K_{Yp}^D}\left(A^T Yp + \left(K_{Yp}^D + Yp\right)\left(Yp + [YpZ] - Y^T\right)\right)}_{c} = 0. \tag{C.49}$$

Only one of the solutions is physically meaningful

$$Y = \frac{1}{2}\left(-b + \sqrt{b^2 - 4c}\right). \tag{C.50}$$

The complex $[YpZ]$ is determined by solving eq. (C.12) in the stationary state for $[YpZ]$ and using the conservation of phosphatases $Z^T = Z + [YpZ]$, yielding

$$[YpZ] = \frac{Z^T Yp}{K_Z^M + Yp}, \quad K_Z^M = \frac{k_Z^2 + k_Z^3}{k_Z^1}. \tag{C.51}$$

By adding eqs. (C.11) + (C.12) we get an equation determining Yp in the stationary state

$$k_Y Y A[P_1 p] + k_Y^5 A[P_2 Yp] - k_Y^4 Ypa[P_2] - k_Z^3 [YpZ] = 0. \tag{C.52}$$

Since we derived expressions giving us the functional dependence on Yp for all other variables, this is the only equation from our previous set of eqs. (C.10) to (C.19) we have to solve numerically.

The equations for CheB are derived in a similar way. We introduce the state variables analogue to eqs. (C.26) and (C.27)

$$A[P_2 B] = A[P_2 B|P_1] + A[P_2 B|P_1 p] \tag{C.53}$$

$$A[P_2 Bp] = A[P_2 Bp|P_1] + A[P_2 Bp|P_1 p], \tag{C.54}$$

and sum up the corresponding eqs. (C.22) + (C.23) and (C.24) + (C.25), respectively:

$$\frac{d}{dt} A[P_2 B] = k_B^1 B A[P_2] - k_B^2 A[P_2 B] - k_B^3 A[P_2 B|P_1 p] \tag{C.55}$$

$$\frac{d}{dt} A[P_2 Bp] = k_B^3 A[P_2 B|P_1 p] - k_B^5 A[P_2 Bp]. \tag{C.56}$$

Similarly to eq. (C.35) we can rewrite eq. (C.23) for $A[P_2 B|P_1 p]$,

$$\frac{d}{dt} A[P_2 B|P_1 p] = k_B^1 B A[P_2|P_1 p] + k_A p_A (A[P_2 B] - A[P_2 B|P_1 p]) - (k_B^2 + k_B^3 + k_Y Y) A[P_2 B|P_1 p]. \tag{C.57}$$

Solving the stationary equations (C.55), (C.56) and (C.57), we get for $A[P_2 Bp]$

$$A[P_2 Bp] = B \Gamma_B \tag{C.58}$$

$$\Gamma_B = \frac{k_B^1 k_B^3 (k_B^2 A[P_2|P_1 p] + k_A p_A A[P_2])}{k_B^5 ((k_B^2 + k_B^3)(k_B^2 + k_A p_A) + k_B^2 k_Y Y)}. \tag{C.59}$$

The equation governing free phosphorylated CheB, eq. (C.21), can be rewritten in terms of $A[P_2 Bp]$ and $A[P_1 p]$

$$\frac{d}{dt} [Bp] = k_B^5 A[P_2 Bp] + k_B B A[P_1 p] - \gamma_B Bp. \tag{C.60}$$

Making the approximation $B \approx B^T - Bp$, we only have to solve eq. (C.60) numerically for Bp in the stationary state.

Rate constants and concentrations

Rate constants are mostly taken from *in vitro* measurements (Dennis Bray's website: http://www.pdn.cam.a
cell/Rates.html) or if unknown set to reasonable values to reflect *in vivo* behavior as given
by FRET measurements of kinase activity.

k_A	$\sim 40\, s^{-1}$	autophosphorylation of CheA
k_Y^1	$\sim 120\, \mu M^{-1} s^{-1}$	CheY to P_2 domain
K_Y^D	$\sim 1.3\, \mu M$	dissociation constant of CheY at P_2 domain
k_Y^3	$\sim 800\, s^{-1}$	phosphotransfer of P_1p to CheY at P_2 domain
k_Y	$\sim 3.2\, \mu M^{-1} s^{-1}$	direct phosphotransfer from P_1p to CheY
k_Y^4	$\sim 120\, \mu M^{-1} s^{-1}$	CheYp to P_2 domain
K_{Yp}^D	$\sim 2.7\, \mu M$	dissociation constant of CheYp at P_2 domain
k_Z^3	$\sim 8\, s^{-1}$	k_{cat} of CheYpCheZ
K_Z^M	$\sim 7.5\, \mu M$	Michaelis-Menten constant of CheYpCheZ
k_B^1	$\sim 0.4\, \mu M^{-1} s^{-1}$	CheB to P_2 domain
K_B^D	$\sim 1\, \mu M$	dissociation constant of CheB at P_2 domain
k_B^3	$\sim 800\, s^{-1}$	phosphotransfer of P_1p to CheB at P_2 domain
k_B	$\sim 0.1\, \mu M^{-1} s^{-1}$	direct phosphotransfer from P_1p to CheB
k_B^4	$\sim 0\, \mu M^{-1} s^{-1}$	CheBp to P_2 domain
k_B^5	$\sim 4\, s^{-1}$	CheBp off P_2 domain
γ_B	$\sim 1\, s^{-1}$	autodephosphorylation of CheBp
k_R	$\sim 0.4\, s^{-1}$	CheR methylation rate of receptors
k_{Bp}	$\sim 17.9\, s^{-1}$	CheBp demethylation rate of receptors
p_A	$\sim 4 \cdot 10^{-2}$	fraction active receptors in the adapted state

Concentrations are calculated using protein copy numbers as given by Li and Hazelbauer (83)
for rich media and assuming an effective cell Volume of $0.7 \mu m^3$.

Bibliography

[1] Alon, U. (2007). *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman & Hall.

[2] Alon, U., Camarena, L., Surette, M. G., Aguera y Arcas, B., Liu, Y., Leibler, S., and Stock, J. B. (1998). Response regulator output in bacterial chemotaxis. *The EMBO journal*, 17(15):4238–48.

[3] Alon, U., Surette, M. G., Barkai, N., and Leibler, S. (1999). Robustness in bacterial chemotaxis. *Nature*, 397(6715):168–71.

[4] Amann, E., Ochs, B., and Abel, K. J. (1988). Tightly regulated tac promoter vectors useful for the expression of unfused and fused proteins in Escherichia coli. *Gene*, 69(2):301–15.

[5] Anand, G. S. and Stock, A. M. (2002). Kinetic basis for the stimulatory effect of phosphorylation on the methylesterase activity of CheB. *Biochemistry*, 41(21):6752–60.

[6] Baker, M. D., Wolanin, P. M., and Stock, J. B. (2006). Signal transduction in bacterial chemotaxis. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 28(1):9–22.

[7] Barkai, N. and Leibler, S. (1997). Robustness in simple biochemical networks. *Nature*, 387(6636):913–7.

[8] Baughman, G. and Nomura, M. (1983). Localization of the target site for translational regulation of the L11 operon and direct evidence for translational coupling in Escherichia coli. *Cell*, 34(3):979–88.

[9] Beard, D. A. and Qian, H. (2008). *Chemical Biophysics: Quantitative Analysis of Cellular Systems*. Cambridge University Press.

[10] Becskei, A. and Serrano, L. (2000). Engineering stability in gene networks by autoregulation. *Nature*, 405(6786):590–3.

[11] Berg, H. C. (1975). Chemotaxis in bacteria. *Annual review of biophysics and bioengineering*, 4(00):119–36.

[12] Berg, H. C. and Brown, D. A. (1972). Chemotaxis in Escherichia coli analysed by three-dimensional tracking. *Nature*, 239(5374):500–4.

[13] Beyer, D., Skripkin, E., Wadzack, J., and Nierhaus, K. H. (1994). How the ribosome moves along the mRNA during protein synthesis. *The Journal of biological chemistry*, 269(48):30713–7.

[14] Blat, Y. and Eisenbach, M. (1994). Phosphorylation-dependent binding of the chemotaxis signal molecule CheY to its phosphatase, CheZ. *Biochemistry*, 33(4):902–6.

[15] Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129(3):897–907.

[16] Burg, K., Haf, H., and Wille, F. (1993). *Höhere Mathematik für Ingenieure Band III*. B.G. Teubner, Stuttgart.

[17] Burg, K., Haf, H., and Wille, F. (1994). *Höhere Mathematik für Ingenieure Band IV*. B.G. Teubner, Stuttgart.

[18] Cai, L., Friedman, N., and Xie, X. S. (2006). Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–62.

[19] Cannarrozzi, G., Schraudolph, N. N., Faty, M., von Rohr, P., Friberg, M. T., Roth, A. C., Gonnet, P., Gonnet, G., and Barral, Y. (2010). A role for codon order in translation dynamics. *Cell*, 141(2):355–67.

[20] Carpentier, A.-S., Torrésani, B., Grossmann, A., and Hénaut, A. (2005). Decoding the nucleoid organisation of Bacillus subtilis and Escherichia coli through gene expression data. *BMC genomics*, 6:84.

[21] Chiaruttini, C., Milet, M., and Springer, M. (1996). A long-range RNA-RNA interaction forms a pseudoknot required for translational control of the IF3-L35-L20 ribosomal protein operon in Escherichia coli. *The EMBO journal*, 15(16):4402–13.

[22] Chiaruttini, C., Milet, M., and Springer, M. (1997). Translational coupling by modulation of feedback repression in the IF3 operon of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 94(17):9208–13.

[23] Clarke, T. F. and Clark, P. L. (2010). Increased incidence of rare codon clusters at 5' and 3' gene termini: implications for function. *BMC genomics*, 11:118.

[24] Cluzel, P., Surette, M., and Leibler, S. (2000). An ultrasensitive bacterial motor revealed by monitoring signaling proteins in single cells. *Science (New York, N.Y.)*, 287(5458):1652–5.

[25] Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–3.

[26] Crick, F. H. (1955). On Degenerate Templates and the Adaptor Hypothesis: A Note for the RNA Tie Club.

[27] Crick, F. H. (1968). The origin of the genetic code. *Journal of Molecular Biology*, 38(3):367–379.

[28] Darty, K., Denise, A., and Ponty, Y. (2009). VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics (Oxford, England)*, 25(15):1974–5.

[29] Davidson, C. J. and Surette, M. G. (2008). Individuality in bacteria. *Annual review of genetics*, 42:253–68.

[30] de Smit, M. H. and van Duin, J. (1990). Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 87(19):7668–72.

[31] de Smit, M. H. and van Duin, J. (1994). Control of translation by mRNA secondary structure in Escherichia coli. A quantitative analysis of literature data. *Journal of molecular biology*, 244(2):144–50.

[32] Dobrzynski, M. and Bruggeman, F. J. (2009). Elongation dynamics shape bursty transcription and translation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(8):2583–8.

[33] Doherty, E. A. and Doudna, J. A. (2000). Ribozyme structures and mechanisms. *Annual review of biochemistry*, 69:597–615.

[34] Dong, H., Nilsson, L., and Kurland, C. G. (1996). Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. *Journal of molecular biology*, 260(5):649–63.

[35] Drummond, D. A. and Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2):341–52.

[36] Ebeling, W., Engel, A., and Feistel, R. (1990). *Physik Der Evolutionsprozesse*. Akad.-Verlag.

[37] Eddy, S. R. (2004). How do RNA folding algorithms work? *Nature biotechnology*, 22(11):1457–8.

[38] Eldar, A. and Elowitz, M. B. (2010). Functional roles for noise in genetic circuits. *Nature*, 467(7312):167–73.

[39] Elf, J., Nilsson, D., Tenson, T., and Ehrenberg, M. (2003). Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science (New York, N.Y.)*, 300(5626):1718–22.

[40] Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science (New York, N.Y.)*, 297(5584):1183–6.

[41] Endres, R. G. and Wingreen, N. S. (2006). Precise adaptation in bacterial chemotaxis through "assistance neighborhoods". *Proceedings of the National Academy of Sciences of the United States of America*, 103(35):13040–4.

[42] Eyre-Walker, A. and Bulmer, M. (1993). Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucleic acids research*, 21(19):4599–603.

[43] Gegner, J. A., Graham, D. R., Roth, A. F., and Dahlquist, F. W. (1992). Assembly of an MCP receptor, CheW, and kinase CheA complex in the bacterial chemotaxis signal transduction pathway. *Cell*, 70(6):975–82.

[44] Gillespie, J. H. (2004). *Population Genetics: A Concise Guide*. JHU Press.

[45] Gingold, H. and Pilpel, Y. (2011). Determinants of translation efficiency and accuracy. *Molecular systems biology*, 7:481.

[46] Golding, I., Paulsson, J., Zawilski, S. M., and Cox, E. C. (2005). Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–36.

[47] Görke, B. and Stülke, J. (2008). Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nature reviews. Microbiology*, 6(8):613–24.

[48] Govantes, F., Andújar, E., and Santero, E. (1998). Mechanism of translational coupling in the nifLA operon of Klebsiella pneumoniae. *The EMBO journal*, 17(8):2368–77.

[49] Grantham, R., Gautier, C., and Gouy, M. (1980). Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic acids research*, 8(9):1893–912.

[50] Gu, W., Zhou, T., and Wilke, C. O. (2010). A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS computational biology*, 6(2):e1000664.

[51] Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., and Altman, S. (1983). The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3 Pt 2):849–57.

[52] Guzman, L. M., Belin, D., Carson, M. J., and Beckwith, J. (1995). Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *Journal of bacteriology*, 177(14):4121–30.

[53] He, L. and Hannon, G. J. (2004). MicroRNAs: small RNAs with a big role in gene regulation. *Nature reviews. Genetics*, 5(7):522–31.

[54] Hershberg, R. and Petrov, D. A. (2008). Selection on codon bias. *Annual review of genetics*, 42(1):287–99.

[55] Herzel, H. and Große, I. (1997). Correlations in DNA sequences: The role of protein coding segments. *Physical Review E*, 55(1):800–810.

[56] Herzel, H., Weiss, O., and Trifonov, E. N. (1999). 10-11 bp periodicities in complete genomes reflect protein structure and DNA folding. *Bioinformatics (Oxford, England)*, 15(3):187–93.

[57] Hess, J. F., Oosawa, K., Kaplan, N., and Simon, M. I. (1988). Phosphorylation of three proteins in the signaling pathway of bacterial chemotaxis. *Cell*, 53(1):79–87.

[58] Hoagland, M. (2004). Enter transfer RNA. *Nature*, 431(7006):249.

[59] Hofacker, I. L. (2009). RNA secondary structure analysis using the Vienna RNA package. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, Chapter 12:Unit12.2.

[60] Ikemura, T. (1981). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *Journal of molecular biology*, 146(1):1–21.

[61] Ingolia, N. T., Brar, G. a., Rouskin, S., McGeachy, A. M., and Weissman, J. S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nature Protocols*, 7(8):1534–1550.

[62] Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., and Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science (New York, N.Y.)*, 324(5924):218–23.

[63] Itzkovitz, S. and Alon, U. (2007). The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome research*, 17(4):405–12.

[64] Jeong, K. S., Ahn, J., and Khodursky, A. B. (2004). Spatial patterns of transcriptional activity in the chromosome of Escherichia coli. *Genome biology*, 5(11):R86.

[65] Karlin, S. (1998). Global dinucleotide signatures and analysis of genomic heterogeneity. *Current opinion in microbiology*, 1(5):598–610.

[66] Karp, P. D., Ouzounis, C. a., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahrén, D., Tsoka, S., Darzentas, N., Kunin, V., and López-Bigas, N. (2005). Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic acids research*, 33(19):6083–9.

[67] Kashtan, N., Noor, E., and Alon, U. (2007). Varying environments can speed up evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 104(34):13711–6.

[68] Kentner, D. and Sourjik, V. (2009). Dynamic map of protein interactions in the Escherichia coli chemotaxis pathway. *Molecular systems biology*, 5(238):238.

[69] Kentner, D., Thiem, S., Hildenbeutel, M., and Sourjik, V. (2006). Determinants of chemoreceptor cluster formation in Escherichia coli. *Molecular microbiology*, 61(2):407–17.

[70] Kertesz, M., Wan, Y., Mazor, E., Rinn, J. L., Nutter, R. C., Chang, H. Y., and Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467(7311):103–7.

[71] Keseler, I. M., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M., Gama-Castro, S., Muñiz Rascado, L., Bonavides-Martinez, C., Paley, S., Krummenacker, M., Altman, T., Kaipa, P., Spaulding, A., Pacheco, J., Latendresse, M., Fulcher, C., Sarker, M., Shearer, A. G., Mackie, A., Paulsen, I., Gunsalus, R. P., and Karp, P. D. (2011). EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic acids research*, 39(Database issue):D583–90.

[72] Keymer, J. E., Endres, R. G., Skoge, M., Meir, Y., and Wingreen, N. S. (2006). Chemosensing in Escherichia coli: two regimes of two-state receptors. *Proceedings of the National Academy of Sciences of the United States of America*, 103(6):1786–91.

[73] Kollmann, M., Løvdok, L., Bartholomé, K., Timmer, J., and Sourjik, V. (2005). Design principles of a bacterial signalling network. *Nature*, 438(7067):504–7.

[74] Komar, A. A. (2009). A pause for thought along the co-translational folding pathway. *Trends in biochemical sciences*, 34(1):16–24.

[75] Kortmann, J. and Narberhaus, F. (2012). Bacterial RNA thermometers: molecular zippers and switches. *Nature reviews. Microbiology*, 10(4):255–65.

[76] Kudla, G. R., Murray, A. W., Tollervey, D., and Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in Escherichia coli. *Science (New York, N.Y.)*, 324(5924):255–8.

[77] Legewie, S., Herzel, H., Westerhoff, H. V., and Blüthgen, N. (2008). Recurrent design patterns in the feedback regulation of the mammalian signalling network. *Molecular systems biology*, 4(190):190.

[78] Lestas, I., Vinnicombe, G., and Paulsson, J. (2010). Fundamental limits on the suppression of molecular fluctuations. *Nature*, 467(7312):174–8.

[79] Levit, M. N., Grebe, T. W., and Stock, J. B. (2002). Organization of the receptor-kinase signaling array that regulates Escherichia coli chemotaxis. *The Journal of biological chemistry*, 277(39):36748–54.

[80] Levit, M. N. and Stock, J. B. (2002). Receptor methylation controls the magnitude of stimulus-response coupling in bacterial chemotaxis. *The Journal of biological chemistry*, 277(39):36760–5.

[81] Li, G.-W., Oh, E., and Weissman, J. S. (2012). The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature*, 484(7395):538–41.

[82] Li, J., Swanson, R. V., Simon, M. I., and Weis, R. M. (1995). The response regulators CheB and CheY exhibit competitive binding to the kinase CheA. *Biochemistry*, 34(45):14626–36.

[83] Li, M. and Hazelbauer, G. L. (2004). Cellular stoichiometry of the components of the chemotaxis signaling complex. *Journal of bacteriology*, 186(12):3687–94.

[84] Liberman, L., Berg, H. C., and Sourjik, V. (2004). Effect of chemoreceptor modification on assembly and activity of the receptor-kinase complex in Escherichia coli. *Journal of bacteriology*, 186(19):6643–6.

[85] Liljeström, P., Laamanen, I., and Palva, E. T. (1988). Structure and expression of the ompB operon, the regulatory locus for the outer membrane porin regulon in Salmonella typhimurium LT-2. *Journal of molecular biology*, 201(4):663–73.

[86] Little, S., Hyde, S., Campbell, C. J., Lilley, R. J., and Robinson, M. K. (1989). Translational coupling in the threonine operon of Escherichia coli K-12. *Journal of bacteriology*, 171(6):3518–22.

[87] Lodish, H., Berk, A., Kaiser, C. A., Krieger, M., Scott, M. P., Bretscher, A., Ploegh, H., and Matsudaira, P. (2007). *Molecular Cell Biology*. W. H. Freeman, 6th edition.

[88] Løvdok, L., Bentele, K., Vladimirov, N., Müller, A., Pop, F. S., Lebiedz, D., Kollmann, M., and Sourjik, V. (2009). Role of translational coupling in robustness of bacterial chemotaxis pathway. *PLoS biology*, 7(8):e1000171.

[89] Løvdok, L., Kollmann, M., and Sourjik, V. (2007). Co-expression of signaling proteins improves robustness of the bacterial chemotaxis pathway. *Journal of biotechnology*, 129(2):173–80.

[90] Lynn, D. J., Singer, G. A. C., and Hickey, D. A. (2002). Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic acids research*, 30(19):4272–7.

[91] Maheshri, N. and O'Shea, E. K. (2007). Living with noisy genes: how cells function reliably with inherent variability in gene expression. *Annual review of biophysics and biomolecular structure*, 36:413–34.

[92] Mao, H., Cremer, P. S., and Manson, M. D. (2003). A sensitive, versatile microfluidic assay for bacterial chemotaxis. *Proceedings of the National Academy of Sciences of the United States of America*, 100(9):5449–54.

[93] Marshall, R. A., Aitken, C. E., Dorywalska, M., and Puglisi, J. D. (2008). Translation at the single-molecule level. *Annual review of biochemistry*, 77:177–203.

[94] Martin, J. A. and Wang, Z. (2011). Next-generation transcriptome assembly. *Nature reviews. Genetics*, 12(10):671–82.

[95] Mathews, D. H. (2006). Revolutions in RNA secondary structure prediction. *Journal of molecular biology*, 359(3):526–32.

[96] Mathews, D. H., Banerjee, A. R., Luan, D. D., Eickbush, T. H., and Turner, D. H. (1997). Secondary structure model of the RNA recognized by the reverse transcriptase from the R2 retrotransposable element. *RNA (New York, N.Y.)*, 3(1):1–16.

[97] Mathews, D. H. and Turner, D. H. (2006). Prediction of RNA secondary structure by free energy minimization. *Current opinion in structural biology*, 16(3):270–8.

[98] McCarthy, J. E. and Bokelmann, C. (1988). Determinants of translational initiation efficiency in the atp operon of Escherichia coli. *Molecular microbiology*, 2(4):455–65.

[99] Mello, B. a. and Tu, Y. (2003). Perfect and near-perfect adaptation in a model of bacterial chemotaxis. *Biophysical journal*, 84(5):2943–56.

[100] Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1):31–46.

[101] Mobilia, M., Reichenbach, T., Hinsch, H., Franosch, T., and Frey, E. (2008). Generic principles of active transport. In *Stochastic Models in Biological Sciences*, Banach Center Publications, pages 101–120, Warsaw. Institute of Mathematics Polish Academy of Sciences.

[102] Moll, I., Grill, S., Gualerzi, C. O., and Bläsi, U. (2002). Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control. *Molecular microbiology*, 43(1):239–46.

[103] Mutoh, N. and Simon, M. I. (1986). Nucleotide sequence corresponding to five chemotaxis genes in Escherichia coli. *Journal of bacteriology*, 165(1):161–6.

[104] NCBI (2012). `http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=tgencodes#SG11`.

[105] Nicolis, G. (1995). *Introduction to Nonlinear Science*. Cambridge University Press.

[106] Nirenberg, M., Caskey, T., Marshall, R., Brimacombe, R., Kellogg, D., Doctor, B., Hatfield, D., Levin, J., Rottman, F., Pestka, S., Wilcox, M., and Anderson, F. (1966). The RNA code and protein synthesis. *Cold Spring Harbor symposia on quantitative biology*, 31:11–24.

[107] Nussinov, R., Pieczenik, G., Griggs, J. R., and Kleitman, D. J. (1978). Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 35(1):68–82.

[108] Onoa, B., Dumont, S., Liphardt, J., Smith, S. B., Tinoco, I., and Bustamante, C. (2003). Identifying kinetic barriers to mechanical unfolding of the T. thermophila ribozyme. *Science (New York, N.Y.)*, 299(5614):1892–5.

[109] Oppenheim, D. S. and Yanofsky, C. (1980). Translational coupling during expression of the tryptophan operon of Escherichia coli. *Genetics*, 95(4):785–95.

[110] Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D., and van Oudenaarden, A. (2002). Regulation of noise in the expression of a single gene. *Nature genetics*, 31(1):69–73.

[111] Papp, B., Pál, C., and Hurst, L. D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature*, 424(6945):194–7.

[112] Parkinson, J. S. and Houts, S. E. (1982). Isolation and behavior of Escherichia coli deletion mutants lacking chemotaxis functions. *Journal of bacteriology*, 151(1):106–13.

[113] Pierce, B. A. (2011). *Genetics: A Conceptual Approac*. W.H. Freeman, 4th edition.

[114] Plotkin, J. B. and Kudla, G. R. (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nature reviews. Genetics*, 12(1):32–42.

[115] Ptashne, M. and Gann, A. (2002). *Genes and Signals*. Cold Spring Harbor Laboratory Press.

[116] Raser, J. M. and O'Shea, E. K. (2004). Control of stochasticity in eukaryotic gene expression. *Science (New York, N.Y.)*, 304(5678):1811–4.

[117] Raser, J. M. and O'Shea, E. K. (2005). Noise in gene expression: origins, consequences, and control. *Science (New York, N.Y.)*, 309(5743):2010–3.

[118] Rausenberger, J., Fleck, C., Timmer, J., and Kollmann, M. (2009). Signatures of gene expression noise in cellular systems. *Progress in biophysics and molecular biology*, 100(1-3):57–66.

[119] Rex, G., Surin, B., Besse, G., Schneppe, B., and McCarthy, J. E. (1994). The mechanism of translational coupling in Escherichia coli. Higher order structure in the atpHA mRNA acts as a conformational switch regulating the access of de novo initiating ribosomes. *The Journal of biological chemistry*, 269(27):18118–27.

[120] Rocha, E. P. C. (2008). The organization of the bacterial genome. *Annual review of genetics*, 42:211–33.

[121] Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S., and Elowitz, M. B. (2005). Gene regulation at the single-cell level. *Science (New York, N.Y.)*, 307(5717):1962–5.

[122] Roulston, M. S. (1999). Estimating the errors on measured entropy and mutual information. *Physica D: Nonlinear Phenomena*, 125(3-4):285–294.

[123] Salgado, H., Moreno-Hagelsieb, G., Smith, T. F., and Collado-Vides, J. (2000). Operons in Escherichia coli: genomic analyses and predictions. *Proceedings of the National Academy of Sciences of the United States of America*, 97(12):6652–7.

[124] Salis, H. M., Mirsky, E. A., and Voigt, C. A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. *Nature biotechnology*, 27(10):946–50.

[125] Schulmeister, S., Ruttorf, M., Thiem, S., Kentner, D., Lebiedz, D., and Sourjik, V. (2008). Protein exchange dynamics at chemoreceptor clusters in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 105(17):6403–8.

[126] Schümperli, D., McKenney, K., Sobieski, D. a., and Rosenberg, M. (1982). Translational coupling at an intercistronic boundary of the Escherichia coli galactose operon. *Cell*, 30(3):865–71.

[127] Segall, J. E., Block, S. M., and Berg, H. C. (1986). Temporal comparisons in bacterial chemotaxis. *Proceedings of the National Academy of Sciences of the United States of America*, 83(23):8987–91.

[128] Serganov, A. and Patel, D. J. (2007). Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. *Nature reviews. Genetics*, 8(10):776–90.

[129] Shah, P. and Gilchrist, M. A. (2011). Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proceedings of the National Academy of Sciences of the United States of America*, 108(25):10231–6.

[130] Shareghi, P., Wang, Y., Malmberg, R., and Cai, L. (2012). Simultaneous prediction of RNA secondary structure and helix coaxial stacking. *BMC genomics*, 13 Suppl 3(Suppl 3):S7.

[131] Sharp, P. M. and Li, W. H. (1987). The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic acids research*, 15(3):1281–95.

[132] Shi, H. and Moore, P. B. (2000). The crystal structure of yeast phenylalanine tRNA at 1.93 A resolution: a classic structure revisited. *RNA (New York, N.Y.)*, 6(8):1091–105.

[133] Shinar, G., Milo, R., Martínez, M. R., and Alon, U. (2007). Input output robustness in simple bacterial signaling systems. *Proceedings of the National Academy of Sciences of the United States of America*, 104(50):19931–5.

[134] Shine, J. and Dalgarno, L. (1974). The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proceedings of the National Academy of Sciences of the United States of America*, 71(4):1342–6.

[135] Singer, G. A. C. and Hickey, D. A. (2003). Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene*, 317(1-2):39–47.

[136] Smith, R. a. and Parkinson, J. S. (1980). Overlapping genes at the cheA locus of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 77(9):5370–4.

[137] Sneppen, K. and Zocchi, G. (2005). *Physics in Molecular Biology*. Cambridge University Press.

[138] Sourjik, V. (2004). Receptor clustering and signal processing in E. coli chemotaxis. *Trends in microbiology*, 12(12):569–76.

[139] Sourjik, V. and Berg, H. C. (2000). Localization of components of the chemotaxis machinery of Escherichia coli using fluorescent protein fusions. *Molecular microbiology*, 37(4):740–51.

[140] Sourjik, V. and Berg, H. C. (2002a). Binding of the Escherichia coli response regulator CheY to its target measured in vivo by fluorescence resonance energy transfer. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12669–74.

[141] Sourjik, V. and Berg, H. C. (2002b). Receptor sensitivity in bacterial chemotaxis. *Proceedings of the National Academy of Sciences of the United States of America*, 99(1):123–7.

[142] Sourjik, V. and Berg, H. C. (2004). Functional interactions between receptors in bacterial chemotaxis. *Nature*, 428(6981):437–41.

[143] Spirin, A. S. (2000). *Ribosomes*. Springer.

[144] Steger, G. (2003). *Bioinformatik: Methoden zur Vorhersage von RNA- und Proteinstrukturen (German Edition)*. Birkhäuser Basel.

[145] Steuer, R., Waldherr, S., Sourjik, V., and Kollmann, M. (2011). Robust signal processing in living cells. *PLoS computational biology*, 7(11):e1002218.

[146] Stewart, R. C. (1997). Kinetic characterization of phosphotransfer between CheA and CheY in the bacterial chemotaxis signal transduction pathway. *Biochemistry*, 36(8):2030–40.

[147] Stewart, R. C., Jahreis, K., and Parkinson, J. S. (2000). Rapid phosphotransfer to CheY from a CheA protein lacking the CheY-binding domain. *Biochemistry*, 39(43):13157–65.

[148] Stewart, R. C. and Van Bruggen, R. (2004). Association and Dissociation Kinetics for CheY Interacting with the P2 Domain of CheA. *Journal of Molecular Biology*, 336(1):287–301.

[149] Storz, G., Vogel, J., and Wassarman, K. M. (2011). Regulation by small RNAs in bacteria: expanding frontiers. *Molecular cell*, 43(6):880–91.

[150] Swain, P. S., Elowitz, M. B., and Siggia, E. D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12795–800.

[151] Tawa, P. and Stewart, R. C. (1994). Kinetics of CheA autophosphorylation and dephosphorylation reactions. *Biochemistry*, 33(25):7917–24.

[152] Thiem, S. and Sourjik, V. (2008). Stochastic assembly of chemoreceptor clusters in Escherichia coli. *Molecular microbiology*, 68(5):1228–36.

[153] Trifonov, E. N. (1987). Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16 S rRNA nucleotide sequences. *Journal of molecular biology*, 194(4):643–52.

[154] Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010a). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, 141(2):344–54.

[155] Tuller, T., Waldman, Y. Y., Kupiec, M., and Ruppin, E. (2010b). Translation efficiency is determined by both codon bias and folding energy. *Proceedings of the National Academy of Sciences of the United States of America*, 107(8):3645–50.

[156] Turner, D. H. (2004). `http://rna.urmc.rochester.edu/NNDB/`.

[157] van Himbergen, J., van Geffen, B., and van Duin, J. (1993). Translational control by a long range RNA-RNA interaction; a basepair substitution analysis. *Nucleic acids research*, 21(8):1713–7.

[158] Veitia, R. a. (2002). Exploring the etiology of haploinsufficiency. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 24(2):175–84.

[159] Wadhams, G. H. and Armitage, J. P. (2004). Making sense of it all: bacterial chemotaxis. *Nature reviews. Molecular cell biology*, 5(12):1024–37.

[160] Wan, Y., Kertesz, M., Spitale, R. C., Segal, E., and Chang, H. Y. (2011). Understanding the transcriptome through RNA structure. *Nature reviews. Genetics*, 12(9):641–55.

[161] Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63.

[162] Warnecke, T. and Hurst, L. D. (2010). GroEL dependency affects codon usage–support for a critical role of misfolding in gene evolution. *Molecular systems biology*, 6(1):340.

[163] Wieser, W. and Krumschnabel, G. (2001). Hierarchies of ATP-consuming processes: direct compared with indirect measurements, and comparative aspects. *The Biochemical journal*, 355(Pt 2):389–95.

[164] Woodson, S. a. (2000). Recent insights on RNA folding mechanisms from catalytic RNA. *Cellular and molecular life sciences : CMLS*, 57(5):796–808.

[165] Xia, T., SantaLucia, J., Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X., Cox, C., and Turner, D. H. (1998). Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37(42):14719–35.

[166] Yi, T. M., Huang, Y., Simon, M. I., and Doyle, J. (2000). Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proceedings of the National Academy of Sciences of the United States of America*, 97(9):4649–53.

[167] Yu, J., Xiao, J., Ren, X., Lao, K., and Xie, X. S. (2006). Probing gene expression in live cells, one protein molecule at a time. *Science (New York, N.Y.)*, 311(5767):1600–3.

[168] Zhang, G., Fedyunin, I., Miekley, O., Valleriani, A., Moura, A., and Ignatova, Z. (2010). Global and local depletion of ternary complex limits translational elongation. *Nucleic acids research*, 38(14):4778–87.

[169] Zhang, G., Hubalewska, M., and Ignatova, Z. (2009). Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature structural & molecular biology*, 16(3):274–80.

[170] Zhang, G. and Ignatova, Z. (2009). Generic algorithm to predict the speed of translational elongation: implications for protein biogenesis. *PloS one*, 4(4):e5036.

[171] Zhao, R., Collins, E. J., Bourret, R. B., and Silversmith, R. E. (2002). Structure and catalytic mechanism of the E. coli chemotaxis phosphatase CheZ. *Nature structural biology*, 9(8):570–5.

[172] Zia, R. K. P., Dong, J. J., and Schmittmann, B. (2011). Modeling Translation in Protein Synthesis with TASEP: A Tutorial and Recent Developments. *Journal of Statistical Physics*, 144(2):405–428.

List of Figures

List of Tables

List of Publications

Linda Løvdok, **Kajetan Bentele**, Nikita Vladimirov, Anette Müller, Ferencz S Pop, Dirk Lebiedz, Markus Kollmann, and Victor Sourjik. Role of translational coupling in robustness of bacterial chemotaxis pathway. *PLoS Biology*, 7(8):e1000171, August 2009

Kajetan Bentele, Martin Falcke. Quasi-steady approximation for ion channel currents. *Biophysical Journal*, 93(8):2597-608, October 2007

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel erarbeitet und verfasst habe.
Ich besitze keinen entsprechenden Doktorgrad und habe mich nicht anderwärts um einen Doktorgrad beworben.
Die dem Promotionsverfahren zugrunde liegende Promotionsordnung ist mir bekannt.

Berlin, den 25.1.2013 Kajetan Bentele